

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**

**REMARKS**

Claims 21-45 are pending in the application. Claims 21, 22, and 32-45 are withdrawn as being drawn to non-elected inventions. Claims 23-31 and 41 are under active consideration. Claim 23 has been amended such that it no longer depends on a non-elected claim 21 in order to address the objection by the Examiner. Claim 31 has been canceled and claim 23 has been amended to remove the embodiment encompassing a polynucleotide encoding a biologically active fragment of a polypeptide having the amino acid sequence of SEQ ID NO:6 in order to address the rejection under 35 U.S.C. § 102. Claim 21 has also been amended to remove the biologically active fragment embodiment in order to expedite prosecution. Claim 30 c) and d) has been amended to recite polynucleotides that are "completely complementary" to address the rejection under 35 U.S.C. § 112, second paragraph. Support for this amendment to claim 30 can be found in the specification, for example, at page 10, lines 19-21. Claim 37 has been amended to replace the term DRASP with HYDRL to properly correspond to the specification (e.g. see page 5, lines 18-21). Entry of these amendments are respectfully requested. Applicants reserve the right to prosecute non-elected subject matter in subsequent divisional applications.

**Comments Regarding Restriction Requirement**

Applicants affirm the election with traverse of Group XX, which corresponds to claims 23-31 and 41 drawn to polynucleotides.

As currently amended, the claims drawn to polypeptides and the claims drawn to polynucleotides are free of prior art. (See the discussion below regarding the rejection under 35 U.S.C. § 102.) Therefore, the objection of lack of unity based on prior art should be withdrawn. Applicants reiterate the request that the Examiner withdraw the Restriction Requirement at least with respect to claims 21, 22, 35, and 36 of Group V, and examine those claims together with the elected polynucleotide claims of Group XX.

The rules under MPEP section 1893.03(d) require the Examiner to apply the Unity of Invention standard PCT Rule 13.2 instead of U.S. restriction/election of species practice in national stage applications, such as the instant application filed under 35 U.S.C. 371. Applicants submit that claims drawn to the polypeptide sequence of SEQ ID NO:6 (*i.e.*, claims 21, 22, 35, and 36) and claims drawn to the elected polynucleotide sequence of SEQ ID NO:22, which

encode SEQ ID NO:6, respectively (*i.e.*, claims 23-31 and 41) meet the unity of invention standards based on the rules concerning unity of invention under the Patent Cooperation Treaty.

The Administrative Instructions Under The Patent Cooperation Treaty, Annex B, Unity of Invention, Part 2, "Examples Concerning Unity of Invention" provide the following guidelines with regard to unity of invention between a protein and the polynucleotide that encodes it:

*Example 17*

Claim 1: Protein X.

Claim 2: DNA sequence encoding protein X.

Expression of the DNA sequence in a host results in the production of a protein which is determined by the DNA sequence. The protein and the DNA sequence exhibit corresponding special technical features. Unity between claims 1 and 2 is accepted.

Applicants submit that Example 17 does apply to the claims of the instant application, since the polynucleotide of SEQ ID NO:22 encodes the polypeptide of SEQ ID NO:6. In particular, claims 21 and 23 meet the unity of invention standards. Claim 23 recites "an isolated polynucleotide encoding a polypeptide of claim 21." Unity of invention is accepted between a protein and the polynucleotide that encodes it. In addition, the recited polynucleotides of claim 30 share chemical and structural features in common with the polynucleotides of claim 23. Refusal to examine claims drawn to the polynucleotides and polypeptides together on the grounds that the polynucleotides are "chemically distinct" from the polypeptides is improper.

**Rejoinder of method claims upon allowance of product claims under U.S. practice**

The Examiner is reminded that claims 32-34, 39, and 40, drawn to methods of using the elected polynucleotides of Group XX should be rejoined per the Commissioner's Notice in the Official Gazette of March 26, 1996, entitled "Guidance on Treatment of Product and Process Claims in light of *In re Ochiai*, *In re Brouwer* and 35 U.S.C. § 103(b)" which sets forth the rules, upon allowance of any product claim, for rejoinder of process claims covering the same scope of products. Applicants request that claims 32-34, 39, and 40 be rejoined and examined upon allowance of any claim drawn to the polynucleotides of Group XX.

**Objection to the Specification**

The Examiner objected to the presence of references to hyperlinks and/or other forms of browser-executable code in the specification (Office Action, page 7). Applicants did not intend

to have active links in the specification, nor to incorporate the subject matter of websites by reference to such hyperlinks. Applicants have amended the specification to remove active hyperlinks and therefore respectfully request that the Examiner withdraw the objection to the specification.

**Objections to the claims**

Claims 23-29 and 41 are objected to because of their dependence from non-elected claims 21 and 22. Claim 23 has been rewritten in independent form as requested by the Examiner; however, as mentioned above, Applicants believe that the claims drawn to the polypeptides of the invention, according to the unity of invention standard, should be examined with the elected claims drawn to the polynucleotides currently under examination. Applicants request reconsideration and believe amending the remainder of the claims at this time would be premature.

**Utility Rejections under 35 U.S.C. §101 and §112, First Paragraph**

Claims 23-31 and 41 are rejected under 35 U.S.C. §§ 101 and 112, first paragraph, based on the allegation that the claimed invention lacks patentable utility. The Office Action alleges in particular that “the claimed invention is not supported by either a specific and substantial asserted utility or well established utility” (Office Action, page 8). Applicants respectfully traverse the rejections.

**The rejection of claims 23-31 and 41 is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well known to one of ordinary skill in the art.**

The invention at issue is a polynucleotide corresponding to a gene that is expressed in gastrointestinal, reproductive, hematopoietic/immune, and cardiovascular tissues. The claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires knowledge of how the polypeptide coded for by the polynucleotide actually functions.

Applicants submit with this response the Declaration of Bedilion describing some of the practical uses of the claimed invention in gene and protein expression monitoring applications.



The Bedilion Declaration demonstrates that the positions and arguments made by the Patent Examiner with respect to the utility of the claimed polynucleotide are without merit.

The Bedilion Declaration describes, in particular, how the claimed expressed polynucleotide can be used in gene expression monitoring applications that were well-known at the time the patent application was filed, and how those applications are useful in developing drugs and monitoring their activity. Dr. Bedilion states that the claimed invention is a useful tool when employed as a highly specific probe in a cDNA microarray:

Persons skilled in the art would appreciate that cDNA microarrays that contained the SEQ ID NO:6-encoding polynucleotides would be a more useful tool than cDNA microarrays that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating cell proliferation, immune system, genetic, and neurological disorders for such purposes as evaluating their efficacy and toxicity. (Bedilion Declaration, ¶ 15.)

The Patent Examiner contends that the claimed polynucleotide cannot be useful without precise knowledge of its biological function, or the biological function of the polypeptide it encodes. But the law has never required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotide in the absence of any knowledge as to the precise function of the protein encoded by it. The uses of the claimed polynucleotide in gene expression monitoring applications are in fact independent of its precise biological function.

## **I. The applicable legal standard**

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed.

Cir. 1992) (“to violate Section 101 the claimed device must be totally incapable of achieving a useful result”); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention “is incapable of serving any beneficial end”).

*Juicy Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: “[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility.” *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. See *Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a “nebulous expression” such as “biological activity” or “biological properties” that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be “substantial.” *Brenner*, 383 U.S. at 534. A “substantial” utility is a practical, “real-world” utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a “well-established” utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examining Procedure at § 706.03(a). Only if there is no “well-established” utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the

Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a “substantial likelihood” of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

**III. Uses of the claimed polynucleotides for diagnosis of conditions and disorders characterized by expression of HYDRL, for toxicology testing, and for drug discovery are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph**

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are “well-established” uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application’s specification. These uses are explained, in detail, in the Bedilion Declaration accompanying this response. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

**A. The use of HYDRL for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer “specific benefits” to the public**

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotide.

The instant application is the National Stage of International Application No. PCT/US99/27009, filed May 8, 2001, which claims the benefit under 35 U.S.C. § 119(e) of provisional application, United States Serial No. 60/135,519, filed on May 21, 1999 (hereinafter “the Hillman ‘519 application”).

In his Declaration, Dr. Bedilion explains the many reasons why a person skilled in the art reading the Hillman ‘519 application on May 21, 1999 would have understood that application to disclose the claimed polynucleotide to be useful for a number of gene expression monitoring

applications, *e.g.*, as a highly specific probe for the expression of that specific polynucleotide in connection with the development of drugs and the monitoring of the activity of such drugs (Bedilion Declaration at, *e.g.*, ¶¶ 10-15). Much, but not all, of Dr. Bedilion's explanation concerns the use of the claimed polynucleotide in cDNA microarrays of the type first developed at Stanford University for evaluating the efficacy and toxicity of drugs, as well as for other applications (Bedilion Declaration at, *e.g.*, ¶¶ 12 and 15).<sup>1</sup>

In connection with his explanations, Dr. Bedilion states that the "Hillman '519 specification would have led a person skilled in the art on May 21, 1999 who was using gene expression monitoring in connection with working on developing new drugs for the treatment of cell proliferation, immune system, genetic, and neurological disorders [a] to conclude that a cDNA microarray that contained the SEQ ID NO:6-encoding polynucleotides would be a highly useful tool, and [b] to request specifically that any cDNA microarray that was being used for such purposes contain the SEQ ID NO:6-encoding polynucleotides" (Bedilion Declaration, ¶ 15). For example, as explained by Dr. Bedilion, "[p]ersons skilled in the art would [have appreciated on May 21, 1999] that a cDNA microarray that contained the SEQ ID NO:6-encoding polynucleotides would be a more useful tool than a cDNA microarray that did not contain the polynucleotides in connection with conducting gene expression monitoring studies on proposed (or actual) drugs for treating cell proliferation, immune system, genetic, and neurological disorders for such purposes as evaluating their efficacy and toxicity." *Id.*

In support of those statements, Dr. Bedilion provided detailed explanations of how cDNA technology can be used to conduct gene expression monitoring evaluations, with extensive citations to pre-May 21, 1999 publications showing the state of the art on May 21, 1999 (Bedilion Declaration, ¶¶ 10-14). While Dr. Bedilion's explanations in paragraph 15 of his Declaration include almost three pages of text and six subparts (a)-(f), he specifically states that his explanations are not "all-inclusive." *Id.* For example, with respect to toxicity evaluations, Dr. Bedilion had earlier explained how persons skilled in the art who were working on drug development on May 21, 1999 (and for several years prior to May 21, 1999) "without any doubt"

---

<sup>1</sup>Dr. Bedilion also explained, for example, why persons skilled in the art would also appreciate, based on the Hillman '519 specification, that the claimed polynucleotide would be useful in connection with developing new drugs using technology, such as Northern analysis, that predated by many years the development of the cDNA technology (Bedilion Declaration, ¶ 16).

appreciated that the toxicity (or lack of toxicity) of any proposed drug was “one of the most important criteria to be evaluated in connection with the development of the drug” and how the teachings of the Hillman ‘519 application clearly include using differential gene expression analyses in toxicity studies (Bedilion Declaration, ¶ 10).

Thus, the Bedilion Declaration establishes that persons skilled in the art reading the Hillman ‘519 application at the time it was filed “would have wanted their cDNA microarray to have a [SEQ ID NO:6-encoding polynucleotide probe] because a microarray that contained such a probe (as compared to one that did not) would provide more useful results in the kind of gene expression monitoring studies using cDNA microarrays that persons skilled in the art have been doing since well prior to May 21, 1999” (Bedilion Declaration, ¶ 15, item (f)). This, by itself, provides more than sufficient reason to compel the conclusion that the Hillman ‘519 application disclosed to persons skilled in the art at the time of its filing substantial, specific and credible real-world utilities for the claimed polynucleotide.

Nowhere does the Patent Examiner address the fact that, as described on pages pp. 35 and 42-43 of the Hillman ‘519 application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays – probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon v. Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 (“Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)” (emphasis added)).

Though Applicants need not so prove to demonstrate utility, there can be no reasonable dispute that persons of ordinary skill in the art have numerous uses for information about relative gene expression including, for example, understanding the effects of a potential drug for treating cell proliferation, immune system, genetic, and neurological disorders. Because the patent application states explicitly that the claimed polynucleotide is known to be expressed both in normal cells as well as cancerous cells (see the Hillman '519 application at Table 3), and expresses a protein that is a member of a class of leucine-rich glycoproteins known to be associated with diseases such as cell proliferation, immune system, genetic, and neurological disorders, there can be no reasonable dispute that a person of ordinary skill in the art could put the claimed invention to such use. In other words, the person of ordinary skill in the art can derive more information about a potential cell proliferation, immune system, genetic, and neurological disorders drug candidate or potential toxin with the claimed invention than without it (see Bedilion Declaration at, e.g., ¶ 15, subparts (e)-(f)).

The Bedilion Declaration shows that a number of pre-May 21, 1999 publications confirm and further establish the utility of cDNA microarrays in a wide range of drug development gene expression monitoring applications at the time the Hillman '519 application was filed (Bedilion Declaration ¶¶ 10-14; Bedilion Exhibits A-G). Indeed, Brown and Shalon U.S. Patent No. 5,807,522 (the Brown '522 patent, Bedilion Exhibit D), which issued from a patent application filed in June 1995 and was effectively published on December 29, 1995 as a result of the publication of a PCT counterpart application, shows that the Patent Office recognizes the patentable utility of the cDNA technology developed in the early to mid-1990s. As explained by Dr. Bedilion, among other things (Bedilion Declaration, ¶ 12):

The Brown '522 patent further teaches that the “[m]icroarrays of immobilized nucleic acid sequences prepared in accordance with the invention” can be used in “numerous” genetic applications, including “monitoring of gene expression” applications (see Bedilion Tab D at col. 14, lines 36-42). The Brown '522 patent teaches (a) monitoring gene expression (i) in different tissue types, (ii) in different disease states, and (iii) in response to different drugs, and (b) that arrays disclosed therein may be used in toxicology studies (see Bedilion Tab D at col. 15, lines 13-18 and 52-58; and col. 18, lines 25-30).

Literature reviews published shortly after the filing of the Hillman '519 application describing the state of the art further confirm the claimed invention's utility. Rockett et al.

confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

\* \* \*

Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

\* \* \*

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original)

Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, Xenobiotica 29:655-691 (July 1999).

In another article, Lashkari et al. state explicitly that sequences that are merely “predicted” to be expressed (predicted Open Reading Frames, or ORFs) – the claimed invention in fact is known to be expressed – have numerous uses:

Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons– they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay. (emphasis added)

Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (Aug. 1997).

**B. The use of polynucleotides coding for polypeptides expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now “well-established”**

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, as described by Bedilion in his Declaration.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., *supra*:

Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Rockett et al., page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, *Molecular Carcinogenesis* 24:153-159 (1999); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, *Toxicology Letters* 112-13:467-471 (2000).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. “Arrays are at their most powerful when they contain the entire genome of the species they are being used to study.” John C. Rockett and David J. Dix, Application of DNA arrays to



toxicology, Environ. Health Perspec.107:681-685 (1999). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding, indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

In fact, the potential benefit to the public, in terms of lives saved and reduced health care costs, are enormous. Evidence of the benefits of this information include:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.
- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.
- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be overturned regardless of their merit.

**C. The similarity of the polypeptide encoded by the claimed invention to another polypeptide of utility demonstrates utility**

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, it is undisputed that the claimed polynucleotide encodes for a protein having the sequence shown as SEQ ID NO:6 in the patent application and referred to as HYDRL-6 in that application. Applicants have demonstrated that HYDRL-6 is a member of the leucine-rich repeat (LRR) family (See specification at Table 2). A recent Blast analysis (Exhibit A) shows that the SEQ ID NO:6 polypeptide is 100% identical to the human leucine-rich  $\alpha$ 2-glycoprotein (g15321646), a member of the LRR family involved in neutrophilic granulocytic differentiation (O'Donnell et al. (2002) J. Leuk. Biol. 72:478-485). This corroborates the statement on page 42 of the Specification that the SEQ ID NO:6 polypeptide and the polynucleotides encoding it may be useful in the diagnosis and treatment of immune disorders.

**D. Objective evidence corroborates the utilities of the claimed invention**

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a “real-world” utility exists. “Real-world” evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Applicants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's databases have proven to be valuable in, for example, the identification and development of drug candidates. Page et al., in discussing the identification and assignment of candidate drug targets, state that "rapid identification and assignment of candidate targets and markers represents a huge challenge ... [t]he process of annotation is similarly aided by the quantity and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals)" Page, M.J. et al., "Proteomics: a major new technology for the drug discovery process," *Drug Discov. Today* 4:55-62 (1999), see page 58, col. 2). As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's invention of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

### **III. The Patent Examiner's rejections are without merit**

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established utilities for the claimed polynucleotide are not "specific, substantial, and credible" utilities. (Office Action at pages 8-11). The Examiner is incorrect both as a matter of law and as a matter of fact.

#### **A. The precise biological role or function of an expressed polynucleotide is not required to demonstrate utility**

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an “identifiable benefit” in presently available form. *Juicy Whip Inc. v. Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the Bedilion Declaration (at, *e.g.*, ¶¶ 10 and 15), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called “throwaway” utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

[T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

#### **B. Membership in a class of useful products can be proof of utility**

Despite the evidence that the claimed polynucleotide encodes a polypeptide in the LRR family, the Examiner refused to impute the utility of the members of the LRR family to HYDRL-6. In the Office Action, the Patent Examiner takes the position that, unless Applicants can

identify which particular biological function within the class of LRR proteins is possessed by HYDRL-6, utility cannot be imputed. To demonstrate utility by membership in the class of LRR proteins, the Examiner would require that all LRR proteins possess a “common” utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. See *Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a “general” class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses HYDRL-6 as if the general class in which it is included is not the LRR family, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these “general classes” may contain a substantial number of useless members, the LRR family does not. The LRR family is sufficiently specific to rule out any reasonable possibility that HYDRL-6 would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the LRR class of proteins has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a “substantial likelihood” that the HYDRL-6 encoded by the claimed polynucleotide is useful. It follows that the claimed polynucleotide also is useful.

Even if the Examiner’s “common utility” criterion were correct – and it is not – the LRR family would meet it. It is undisputed that known members of the LRR family closely related to the SEQ ID NO:6 polypeptide are involved in the immune response and cell adhesion (O’Donnell et al., *supra* and Saito et al. (2002) J. Immunol. 168:1050-1059). A person of

ordinary skill in the art need not know any more about how the claimed invention functions to use it, and the Examiner presents no evidence to the contrary. Instead, the Examiner makes the conclusory observation that a person of ordinary skill in the art would need to confirm the activity of any given LRR protein (Office Action, page 9). The Examiner then goes on to assume that the only use for HYDRL-6 absent knowledge as to how the leucine-rich glycoprotein actually works is further study of HYDRL-6 itself.

Not so. As demonstrated by Applicants, knowledge that HYDRL-6 is a leucine-rich glycoprotein is more than sufficient to make it useful for the diagnosis and treatment of cell proliferation, immune system, genetic, and neurological disorders. Indeed, HYDRL-6 has been shown to be expressed in gastrointestinal, reproductive, hematopoietic/immune, and cardiovascular tissues and in tissues associated with cancer or inflammation. (Hillman '519 application at Table 3). The Examiner must accept these facts to be true unless the Examiner can provide evidence or sound scientific reasoning to the contrary. But the Examiner has not done so.

**C. Because the uses of polynucleotides encoding HYDRL in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the invention itself, the claimed invention has substantial utility**

The PTO rejected the claims at issue on the ground that the use of an invention as a tool for research is not a “substantial” use. Because the PTO’s rejection assumes a substantial overstatement of the law, and is incorrect in fact, it must be overturned.

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office has recognized that just because an invention is used in a research setting does not mean that it lacks utility (Section § 2107.01 of the Manual of Patent Examining Procedure, 8<sup>th</sup> Edition, August 2001, under the heading I. Specific and Substantial Requirements, Research Tools):

Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact “useful” in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm.

The Patent Office's actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases. These are acknowledged by the PTO's Training Materials themselves to be useful, as well as DNA sequences used, for example, as markers.

Only a limited subset of research uses are not "substantial" utilities: those in which the only known use for the claimed invention is to be an **object** of further study, thus merely inviting further research. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the only known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945 ("What Applicants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines."). Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other beneficial use in research.

As used in toxicology testing, drug discovery, and disease diagnosis, the claimed invention has a beneficial use in research other than studying the claimed invention or its protein products. It is a tool, rather than an object, of research. The data generated in gene expression monitoring using the claimed invention as a tool is **not** used merely to study the claimed polynucleotide itself, but rather to study properties of tissues, cells, and potential drug candidates and toxins. Without the claimed invention, the information regarding the properties of tissues, cells, drug candidates and toxins is less complete. (Bedilion Declaration at ¶ 15.)

The claimed invention has numerous additional uses as a research tool, each of which alone is a "substantial utility." These include uses such as diagnostic assays (e.g., pages 41-47), chromosomal markers (e.g., page 46), ligand screening assays (e.g., page 33 and 46), and drug screening (pages 46-47).

**IV. By requiring the patent applicant to assert a particular or unique utility, the Patent Examination Utility Guidelines and Training Materials applied by the Patent Examiner misstate the law**

There is an additional, independent reason to overturn the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427,

December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website [www.uspto.gov](http://www.uspto.gov), March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: “specific” utilities which meet the statutory requirements, and “general” utilities which do not. The Training Materials define a “specific utility” as follows:

A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as “gene probe” or “chromosome marker” would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between “specific” and “general” utilities by assessing whether the asserted utility is sufficiently “particular,” *i.e.*, unique (Training Materials at page 52) as compared to the “broad class of invention.” (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) (“With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.”)).

Such “unique” or “particular” utilities never have been required by the law. To meet the utility requirement, the invention need only be “practically useful,” *Natta*, 480 F.2d 1 at 1397, and confer a “specific benefit” on the public. *Brenner*, 383 U.S. at 534. Thus, incredible “throw-away” utilities, such as trying to “patent a transgenic mouse by saying it makes great snake food,” do not meet this standard. Karen Hall, Genomic Warfare, *The American Lawyer* 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where “specific utility” is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be “definite,” not particular. *Montedison*, 664 F.2d at 375. Applicants are not aware of any court that has rejected an assertion of utility on the grounds that it is not “particular” or “unique” to the



specific invention. Where courts have found utility to be too “general,” it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had “useful biological activity” was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a “particular” type of cancer was determined to satisfy the specificity requirement). “Particularity” is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Supra* § II.B.2 (*Montedison*, 664 F.2d at 374-75).

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of “general” utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. See *supra* § II.B. Thus the Training Materials cannot be applied consistently with the law.

**V. To the extent the rejection of the claimed invention under 35 U.S.C. § 112, first paragraph, is based on the improper rejection for lack of utility under 35 U.S.C. § 101, it must be reversed.**

The rejection set forth in the Office Action is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under 35 U.S.C. § 112, first paragraph, is based on the improper allegation of lack of patentable utility under 35 U.S.C. § 101, it fails for the same reasons.

CONCLUSION

Applicants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of "lack of specificity," as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, "like a nose of wax,"<sup>2</sup> to target rejections of claims to polypeptide and polynucleotide sequences, as well as to claims to methods of detecting said polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be reversed.

Moreover, to the extent the above rejections were based on the Revised Interim and final Examination Guidelines and Training Materials, those portions of the Guidelines and Training Materials that form the basis for the rejections should be determined to be inconsistent with the law.

Rejections under 35 U.S.C. § 112, second paragraph

Claims 23, 26-28, 30, and 41 are rejected under 35 U.S.C. § 112, second paragraph, as allegedly being indefinite (Office Action, page 11). In particular, it is asserted that in claim 23, the term, "biologically active" is indefinite because "the scope of activities encompassed by this term is vague." In claim 30, the term "complementary" is allegedly indefinite because "it is unclear as to whether the complementary polynucleotides are partial or complete complements" (Office Action, page 11).

---

<sup>2</sup>"The concept of patentable subject matter under §101 is not 'like a nose of wax which may be turned and twisted in any direction \* \* \*.' *White v. Dunbar*, 119 U.S. 47, 51." (*Parker v. Flook*, 198 USPQ 193 (US SupCt 1978))

The biologically active fragments of claim 23 have been canceled; therefore, the rejection with respect to this claim and dependent claims 26-28 and 41 is moot. To expedite prosecution, claim 30 c) and d) has been amended as suggested by the Examiner to recite that the claimed polynucleotides are completely complementary.

For at least the above reasons, withdrawal of the rejection under 35 U.S.C. § 112, second paragraph, is respectfully requested.

**Written description rejections under 35 U.S.C. § 112, first paragraph**

Claims 23, 26-28, 30, 31, and 41 have been rejected under the first paragraph of 35 U.S.C. 112 for alleged lack of an adequate written description. This rejection is respectfully traversed.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*. The invention is, for purposes of the "written description" inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that :

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met. (footnotes omitted.)

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:6 and SEQ ID NO:22 are specifically disclosed in the application (see, for example, page 5, lines 17-25 and page 6, lines 12-14). Variants of SEQ ID NO:22 are described, for example, at page 6, lines 14-16. Incyte clones in which the nucleic acids encoding SEQ ID NO:6 were first identified and libraries from which those clones were isolated are described, for example, at Table 1 of the Specification. Chemical and structural features of SEQ ID NO:6 are described, for example, at Table 2. Given SEQ ID NO:6 and SEQ ID NO:22, one of ordinary skill in the art would recognize naturally occurring variants of SEQ ID NO:22 having 90% sequence identity to SEQ ID NO:22 or polynucleotides encoding a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:6. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

The Office Action has further asserted that the claims are not supported by an adequate written description because "the disclosure of the single representative species of SEQ ID NO:22 is insufficient to be representative of the attributes and features of all species encompassed by the claimed genus of polynucleotides" (Office Action, page 13).

Such a position is believed to present a misapplication of the law.

**1. The present claims specifically define the claimed genus through the recitation of chemical structure**

Court cases in which "DNA claims" have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; i.e., "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides in terms of chemical structure, rather than functional characteristics. For example, the "variant language" of independent claim 30 recites chemical structure to define the claimed genus:

An isolated polynucleotide selected from the group consisting of:...

- b) a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical to a polynucleotide sequence of SEQ ID NO:22...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:6 and SEQ ID NO:22. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides. The polynucleotides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*.

**2. The present claims do not define a genus which is "highly variant"**

Furthermore, the claims at issue do not describe a genus which could be characterized as "highly variant." Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the enclosed reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to leucine-rich glycoprotein proteins related to the amino acid sequence of SEQ ID NO:6. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as leucine-rich glycoprotein proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:6. The "variant language" of the present claims recites, for example, polynucleotides encoding "a naturally occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:6" (note that SEQ ID NO:6 has 347 amino acid residues). This variation is far less than that of all potential leucine-rich glycoprotein proteins related to SEQ ID NO:6, i.e., those leucine-rich glycoprotein proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:6.

**3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications**

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of May 21, 1999. Much has happened in the development of recombinant DNA technology in the 20 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:6 and SEQ ID NO:22, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

#### 4. Summary

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:6 or SEQ ID NO:22. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides defined by the present claims is adequately described, as evidenced by Brenner et al and consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

#### Enablement rejections under 35 U.S.C. § 112, first paragraph

Claims 23, 26-28, 30, 31, and 41 are rejected for allegedly failing to meet the requirements of 35 U.S.C. § 112, first paragraph, on the grounds that the Specification does not provide an enabling disclosure commensurate in scope with the claims (Office Action, page 12). In particular, the Examiner alleges that "the specification, while being enabling for a polynucleotide encoding SEQ ID NO:6 including SEQ ID NO:22, does not reasonably provide enablement for the broad scope of claimed polynucleotides" (Office Action, page 14). Applicants traverse the rejection for at least the following reasons.

As set forth in *In re Marzocchi*, 169 USPQ 367, 369 (CCPA 1971):

The first paragraph of § 112 requires nothing more than **objective enablement**. How such a teaching is set forth, either by the use of illustrative examples or by broad terminology, is of no importance.

As a matter of Patent Office practice, then, a specification disclosure which contains a teaching of the manner and process of making and using the invention in terms which correspond in scope to those used in describing and defining the subject matter sought to be patented *must* be taken as in compliance with the enabling requirement of the first paragraph of § 112 *unless* there is reason to doubt the objective truth of the statements contained therein which must be relied on for enabling support.



Applicants submit that the disclosure amply enables the claimed invention. Given the sequences of SEQ ID NO:6 and SEQ ID NO:22, one of ordinary skill in the art could readily identify a polynucleotide encoding a polypeptide comprising a naturally occurring amino acid sequence at least 90% identical to an amino acid sequence of SEQ ID NO:6 or a polynucleotide comprising a naturally occurring polynucleotide sequence at least 90% identical to a polynucleotide sequence of SEQ ID NO:22, using well known methods of sequence analysis without any undue experimentation. For example, the identification of relevant polynucleotides could be performed by hybridization and/or PCR techniques that were well-known to those skilled in the art at the time the subject application was filed and/or described throughout the Specification of the instant application. See, e.g., page 29, lines 9-20; page 41, line 26 through page 42, line 9; and Example VII at page 53. Thus, one skilled in the art need not make and test vast numbers of polynucleotides. Instead, one skilled in the art need only screen a cDNA library or use appropriate PCR conditions to identify relevant polynucleotides that already exist in nature. The skilled artisan would also know how to use the claimed polynucleotides, for example in expression profiling, disease diagnosis, or detection of related sequences as discussed above. The specification also describes the expression vectors into which the claimed variants and fragments could be inserted, and the construction of fusion proteins (pages 26-30 and Example X at pages 54-55).

The Examiner has cited some references (Branden et al. "Introduction to Protein Structure," Garland Publishing Inc., New York, 1991; Witkowski et al. (1999) Biochemistry 38:11643-11650; Brenner et al. (1999) Trends Genet 15:132-133); and Scott et al. (1999) Nat. Genet. 21:440-443) that supposedly underscore the "unpredictability" of protein chemistry. Again, Applicants respectfully point out that the claims of the instant application are drawn to **naturally occurring** variants. Thus it is not necessary to screen every conceivable variant which might be made using recombinant methods, as all that is claimed are those variant sequences which are found in nature. Through the process of natural selection, nature will have determined the appropriate sequences.

Furthermore, the claims are directed to polynucleotides, not polypeptides, and it is the functionality of the claimed polynucleotides, not the polypeptides encoded by them, that is relevant. Members of the claimed genus of variants may include, for example, mutant alleles associated with diseases, or single nucleotide polymorphisms (SNPs). Members of the claimed

genus of variants may be useful even if they encode defective HYDRL polypeptides. For example, the variant polynucleotides could be used for the detection of sequences related to HYDRL (see the specification, for example, at page 41, lines 26-32) including HYDRL variants that may be associated with disease states, such as the diseases listed on page 42, line 11 through page 44, line 6, of the specification. See the specification at, for example, pages 41-47 for disclosure of how to use the claimed sequences in diagnostic assays.

Further, the Examiner requires working examples (Office Action, page 15). There is no such requirement under the law to provide “working examples.” As set forth in *In re Borkowski*, 164 USPQ 642, 645 (CCPA 1970) (footnote omitted):

However, as we have stated in a number of opinions, a specification need not contain a working example if the invention is otherwise disclosed in such a manner that one skilled in the art will be able to practice it without an undue amount of experimentation.

See also M.P.E.P. 2164.02 as follows:

Compliance with the enablement requirement of 35 U.S.C. 112, first paragraph, does not turn on whether an example is disclosed. An example may be “working” or “prophetic”... A prophetic example describes an embodiment of the invention based on predicted results rather than work actually conducted or results actually achieved.

Thus, there is no requirement under the law to provide “working examples” of what is claimed. Rather, one looks to whether the specification provides a description of how to make what is claimed. The present specification provides the requisite description.

Contrary to the standard set forth in *Marzocchi* and *Borkowski*, the Examiner has failed to provide any *reasons* why one would doubt that the guidance provided by the present specification would enable one to make and use the recited polynucleotides. Hence, a *prima facie* case for non-enablement has not been established. For at least the above reasons, withdrawal of the enablement rejections under 35 U.S.C. § 112, first paragraph, is respectfully requested.

### **Rejection under 35 U.S.C. § 102**

Claims 23, 26-28, 31, and 41 are rejected under 35 U.S.C. § 102(a) as allegedly being anticipated by Jacobs et al. (WO 98/45437) and claim 31 is rejected under 35 U.S.C. § 102(b) as allegedly being anticipated by GenBank Accession No. AA622495 (Office Action at pages 18-19).

Claim 31 has been canceled; therefore, the rejection with respect to this claim is moot.

Claim 23 has been amended to remove the embodiment reciting a polynucleotide encoding a biologically active fragment of a polypeptide having the amino acid sequence of SEQ ID NO:6.

This amendment to claim 23 renders this rejection moot. Therefore, reconsideration and withdrawal of this rejection are respectfully requested.

**CONCLUSION**

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding objections/rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact the undersigned at the number listed below.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108**.

Respectfully submitted,

INCYTE CORPORATION

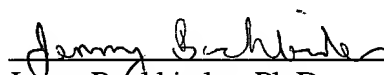
Date: December 19, 2003

  
James M. Verna, Ph.D.

Reg. No. 33,287

Direct Dial Telephone: (650) 845 -5415

Date: December 19, 2003

  
Jenny Buchbinder, Ph.D.

Reg. No. 48,588

Direct Dial Telephone: (650) 843-7212

**Customer No.: 27904**

3160 Porter Drive

Palo Alto, California 94304

Phone: (650) 855-0555

Fax: (650) 849-8886

Enclosures:

1. Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, Xenobiotica 29:655-691 (1999).
2. Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (1997).
3. Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999).

4. Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).
5. John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, 107 Environ. Health Perspec. 107:681-685 (1999).
6. Email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.
7. Brenner et al., Proc. Natl. Acad. Sci. 95:6073-6078 (1998).
8. Page, M.J. et al., Proteomics: a major new technology for the drug discovery process, Drug Discov. Today 4:55-62 (1999).
9. O'Donnell et al. J. Leuk. Biol. 72:478-485 (2002).
10. Saito et al. J. Immunol. 168:1050-1059 (2002).
11. Exhibit A



## Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential

JOHN C. ROCKETT†, DAVID J. ESDAILE‡  
and G. GORDON GIBSON\*

Molecular Toxicology Laboratory, School of Biological Sciences, University of Surrey,  
Guildford, Surrey, GU2 5XH, UK

Received January 8, 1999

1. An important feature of the work of many molecular biologists is identifying which genes are switched on and off in a cell under different environmental conditions or subsequent to xenobiotic challenge. Such information has many uses, including the deciphering of molecular pathways and facilitating the development of new experimental and diagnostic procedures. However, the student of gene hunting should be forgiven for perhaps becoming confused by the mountain of information available as there appears to be almost as many methods of discovering differentially expressed genes as there are research groups using the technique.

2. The aim of this review was to clarify the main methods of differential gene expression analysis and the mechanistic principles underlying them. Also included is a discussion on some of the practical aspects of using this technique. Emphasis is placed on the so-called 'open' systems, which require no prior knowledge of the genes contained within the study model. Whilst these will eventually be replaced by 'closed' systems in the study of human, mouse and other commonly studied laboratory animals, they will remain a powerful tool for those examining less fashionable models.

3. The use of suppression-PCR subtractive hybridization is exemplified in the identification of up- and down-regulated genes in rat liver following exposure to phenobarbital, a well-known inducer of the drug metabolizing enzymes.

4. Differential gene display provides a coherent platform for building libraries and microchip arrays of 'gene fingerprints' characteristic of known enzyme inducers and xenobiotic toxicants, which may be interrogated subsequently for the identification and characterization of xenobiotics of unknown biological properties.

### Introduction

It is now apparent that the development of almost all cancers and many non-neoplastic diseases are accompanied by altered gene expression in the affected cells compared to their normal state (Hunter 1991, Wynford-Thomas 1991, Vogelstein and Kinzler 1993, Semenza 1994, Cassidy 1995, Kleinjan and Van Hegningen 1998). Such changes also occur in response to external stimuli such as pathogenic micro-organisms (Rohn *et al.* 1996, Singh *et al.* 1997, Griffin and Krishna 1998, Lunney 1998) and xenobiotics (Sewall *et al.* 1995, Dogra *et al.* 1998, Ramana and Kohli 1998), as well as during the development of undifferentiated cells (Hecht 1998, Rudin and Thompson 1998, Schneider-Maunoury *et al.* 1998). The potential medical and therapeutic benefits of understanding the molecular changes which occur in any given cell in progressing from the normal to the 'altered' state are enormous. Such profiling essentially provides a 'fingerprint' of each step of a

\* Author for correspondence; e-mail: g.gibson@surrey.ac.uk

† Current Address: US Environmental Protection Agency, National Health and Environmental Effects, Research Laboratory, Reproductive Toxicology Division, Research Triangle Park, NC 27711, USA.

‡ Rhone-Poulenc Agrochemicals, Toxicology Department, Sophia-Antipolis, Nice, France.

altered expression in cells of one population compared to another. These methods have been used to identify differential gene expression in many situations, including invading pathogenic microbes (Zhao *et al.* 1998), in cells responding to extracellular and intracellular microbial invasion (Duguid and Dinanuer 1990, Ragno *et al.* 1997, Maldarelli *et al.* 1998), in chemically treated cells (Syed *et al.* 1997, Rockett *et al.* 1999), neoplastic cells (Liang *et al.* 1992, Chang and Terzaghi-Howe 1998), activated cells (Gurskaya *et al.* 1996, Wan *et al.* 1996), differentiated cells (Hara *et al.* 1991, Guimaraes *et al.* 1995a, b), and different cell types (Davis *et al.* 1984, Hedrick *et al.* 1984, Xhu *et al.* 1998). Although differential expression analysis technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

The field of differential expression analysis is a large and complex one, with many techniques available to the potential user. These can be categorized into several methodological approaches, including:

- (1) Differential screening,
- (2) Subtractive hybridization (SH) (includes methods such as chemical cross-linking subtraction—CCLS, suppression-PCR subtractive hybridization—SSH, and representational difference analysis—RDA),
- (3) Differential display (DD),
- (4) Restriction endonuclease facilitated analysis (including serial analysis of gene expression—SAGE—and gene expression fingerprinting—GEF),
- (5) Gene expression arrays, and
- (6) Expressed sequence tag (EST) analysis.

The above approaches have been used successfully to isolate differentially expressed genes in different model systems. However, each method has its own subtle (and sometimes not so subtle) characteristics which incur various advantages and disadvantages. Accordingly, it is the purpose of this review to clarify the mechanistic principles underlying the main differential expression methods and to highlight some of the broader considerations and implications of this very powerful and increasingly popular technique. Specifically, we will concentrate on the so-called 'open' systems, namely those which do not require any knowledge of gene sequences and, therefore, are useful for isolating unknown genes. Two 'closed' systems (those utilising previously identified gene sequences), EST analysis and the use of DNA arrays, will also be considered briefly for completeness. Whilst emphasis will often be placed on suppression PCR subtractive hybridization (SSH, the approach employed in this laboratory), it is the aim of the authors to highlight, wherever possible, those areas of common interest to those who use, or intend to use, differential gene expression analysis.

### Differential cDNA library screening (DS)

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years. One of the original approaches used to identify such genes was described 20 years ago by St John and Davis (1979). These authors developed a method, termed 'differential plaque filter

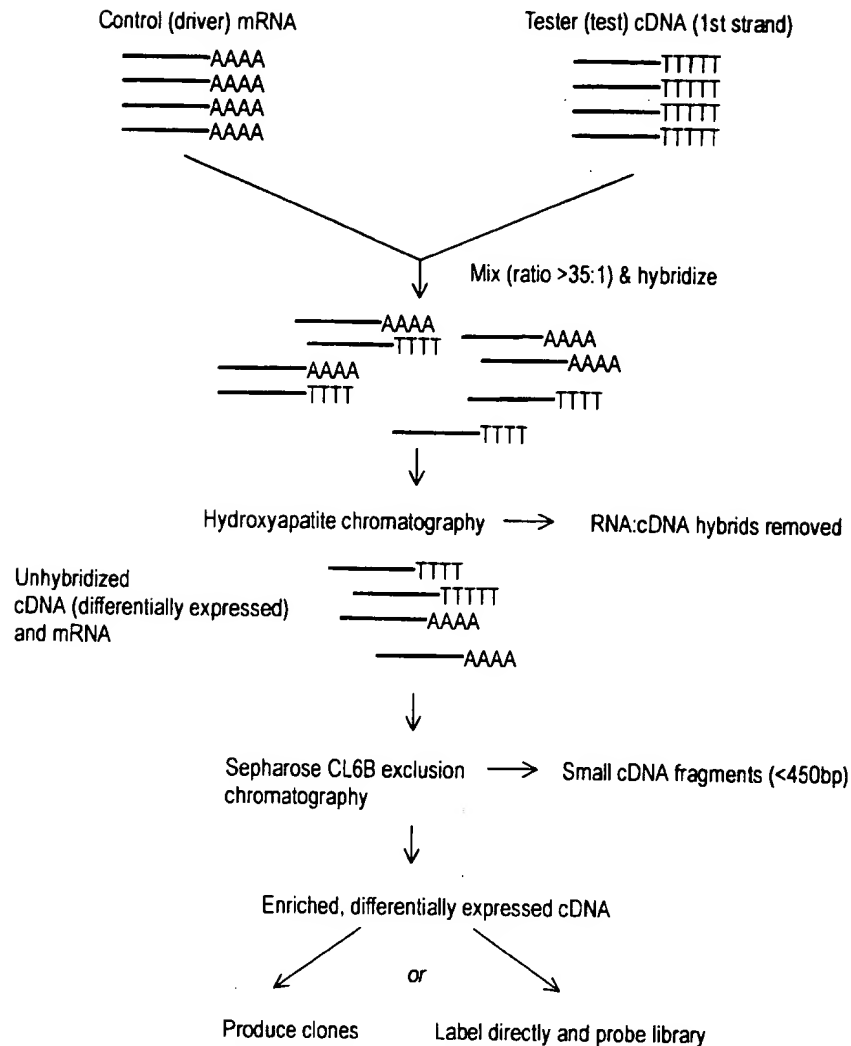


Figure 1. The hydroxyapatite method of subtractive hybridization. cDNA derived from the treated/alterd (tester) population is mixed with a large excess of mRNA from the control (driver) population. Following hybridization, mRNA-cDNA hybrids are removed by hydroxyapatite chromatography. The only cDNAs which remain are those which are differentially expressed in the treated/alterd population. In order to facilitate the recovery of full length clones, small cDNA fragments are removed by exclusion chromatography. The remaining cDNAs are then cloned into a vector for sequencing, or labelled and used directly to probe a library, as described by Sargent and Dawid (1983).

containing a restriction site) ligated to both sides. Both populations are then amplified by PCR, but the driver cDNA population is subsequently digested with the adaptor-containing restriction endonuclease. This serves to cleave the oligo-vector and reduce the amplification potential of the control population. The digested control population is then biotinylated and an excess mixed with tester cDNA. Following denaturation and hybridization, the mix is applied to a biocytin column (streptavidin may also be used) to remove the control population, including heteroduplexes formed by annealing of common sequences from the tester population. The procedure is repeated several times following the addition of fresh



control cDNA. In order to further enrich those species differentially expressed in the tester cDNA, the subtracted tester population is amplified by PCR following every second subtraction cycle. After six cycles of subtraction (three reamplification steps) the reaction mix is ligated into a vector for further analysis.

In a slightly different approach, Hara *et al.* (1991) utilized a method whereby oligo(dT<sub>30</sub>) primers attached to a latex substrate are used to first capture mRNA extracted from the control population. Following 1st strand cDNA synthesis, the RNA strand of the heteroduplexes is removed by heat denaturation and centrifugation (the cDNA-oligotex-dT<sub>30</sub> forms a pellet and the supernatant is removed). A quantity of tester mRNA is then repeatedly hybridized to the immobilized control (driver) cDNA (which is present in 20-fold excess). After several rounds of hybridization the only mRNA molecules left in the tester mRNA population are those which are not found in the driver cDNA-oligotex-dT<sub>30</sub> population. These tester-specific mRNA species are then converted to cDNA and, following the addition of adaptor sequences, amplified by PCR. The PCR products are then ligated into a vector for further analysis using restriction sites incorporated into the PCR primers. A schematic illustration of this subtraction process is shown in figure 2.

However, all these methods utilising physical separation have been described as inefficient due to the requirement for large starting amounts of mRNA, significant loss of material during the separation process and a need for several rounds of hybridization. Hence, new methods of differential expression analysis have recently been designed to eliminate these problems.

#### *Chemical Cross-Linking Subtraction (CCLS)*

In this technique, originally described by Hampson *et al.* (1992), driver mRNA is mixed with tester cDNA (1st strand only) in a ratio of > 20:1. The common sequences form cDNA:mRNA hybrids, leaving the tester specific species as single stranded cDNA. Instead of physically separating these hybrids, they are inactivated chemically using 2,5 diaziridinyl-1,4-benzoquinone (DZQ). Labelled probes are then synthesized from the remaining single stranded cDNA species (unreacted mRNA species remaining from the driver are not converted into probe material due to specificity of Sequenase T7 DNA polymerase used to make the probe) and used to screen a cDNA library made from the tester cell population. A schematic diagram of the system is shown in figure 3.

It has been shown that the differentially expressed sequences can be enriched at least 300-fold with one round of subtraction (Hampson *et al.* 1992), and that the technique should allow isolation of cDNAs derived from transcripts that are present at less than 50 copies per cell. This equates to genes at the low end of intermediate abundance (see table 1). The main advantages of the CCLS approach are that it is rapid, technically simple and also produces fewer false positives than other differential expression analysis methods. However, like the physical separation protocols, a major drawback with CCLS is the large amount of starting material required (at least 10 µg RNA). Consequently, the technique has recently been refined so that a renewable source of RNA can be generated. The degenerate random oligonucleotide primed (DROP) adaptation (Hampson *et al.* 1996, Hampson and Hampson 1997) uses random hexanucleotide sequences to prime solid phase-synthesized cDNA. Since each primer includes a T7 polymerase promoter sequence

at the 5' end, the final pool of random cDNA fragments is a PCR-renewable cDNA population which is representative of the expressed gene pool and can be used to synthesize sense RNA for use as driver material. Furthermore, if the final pool of random cDNA fragments is reamplified using biotinylated T7 primer and random hexamer, the product can be captured with streptavidin beads and the antisense strand eluted for use as tester. Since both target and driver can be generated from the same DROP product, subtraction can be performed in both directions (i.e. for up- and down-regulated species) between two different DROP products.

#### *Representational Difference Analysis (RDA)*

RDA of cDNA (Hubank and Schatz 1994) is an extension of the technique originally applied to genomic DNA as a means of identifying differences between two complex genomes (Lisitsyn *et al.* 1993). It is a process of subtraction and amplification involving subtractive hybridization of the tester in the presence of excess driver. Sequences in the tester that have homologues in the driver are rendered unamplifiable, whereas those genes expressed only in the tester retain the ability to be amplified by PCR. The procedure is shown schematically in figure 4.

In essence, the driver and tester mRNA populations are first converted to cDNA and amplified by PCR following the ligation of an adaptor. The adaptors are then removed from both populations and a new (different) adaptor ligated to the amplified tester population only. Driver and tester populations are next melted and hybridized together in a ratio of 100:1. Following hybridization, only tester:tester homohybrids have 5' adaptors at each end of the DNA duplex and can, thus, be filled in at both 3' ends. Hence, only these molecules are amplified exponentially during the subsequent PCR step. Although tester:driver heterohybrids are present, they only amplify in a linear fashion, since the strand derived from the driver has no adaptor to which the primer can bind. Driver:driver heterohybrids have no adaptors and, therefore, are not amplified. Single stranded molecules are digested with mung bean nuclease before a further PCR-enrichment of the tester:tester homohybrids. The adaptors on the amplified tester population are then replaced and the whole process repeated a further two or three times using an increasing excess of driver (Hubank and Schatz used a tester:driver ratio of 1:400, 1:80000 and 1:800000 for the second, third and fourth hybridizations, respectively). Different adaptors are ligated to the tester between successive rounds of hybridization and amplification to prevent the accumulation of PCR products that might interfere with subsequent amplifications. The final display is a series of differentially expressed gene products easily observable on an ethidium bromide gel.

The main advantages of RDA are that it offers a reproducible and sensitive approach to the analysis of differentially expressed genes. Hubank and Schatz (1994) reported that they were able to isolate genes that were differentially expressed in substantially less than 1% of the cells from which the tester is derived. Perhaps the main drawback is that multiple rounds of ligation, hybridization, amplification and digestion are required. The procedure is, therefore, lengthier than many other differential display approaches and provides more opportunity for operator-induced error to occur. Although the generation of false positives has been noted, this has been solved to some degree by O'Neill and Sinclair (1997) through the use of HPLC-purified adaptors. These are free of the truncated adaptors which appear to be a major source of the false positive bands. A very similar technique to RDA, termed linker capture subtraction (LCS) was described by Yang and Sytowski (1996).

*Suppression PCR Subtractive Hybridization (SSH)*

The most recent adaptation of the SH approach to differential expression analysis was first described by Diatchenko *et al.* (1996) and Gurskaya *et al.* (1996). They reported that a 1000–5000 fold enrichment of rare cDNAs (equivalent to isolating mRNAs present at only a few copies per cell) can be obtained without the need for multiple hybridizations/subtractions. Instead of physical or chemical removal of the common sequences, a PCR-based suppression system is used (see figure 5).

In SSH, excess driver cDNA is added to two portions of the tester cDNA which have been ligated with different adaptors. A first round of hybridization serves to enrich differentially expressed genes and equalize rare and abundant messages. Equalization occurs since reannealing is more rapid for abundant molecules than for rarer molecules due to the second order kinetics of hybridization (James and Higgins 1985). The two primary hybridization mixes are then mixed together in the presence of excess driver and allowed to hybridize further. This step permits the annealing of single stranded complementary sequences which did not hybridize in the primary hybridization, and in doing so generates templates for PCR amplification. Although there are several possible combinations of the single stranded molecules present in the secondary hybridization mix, only one particular combination (differentially expressed in the tester cDNA composed of complimentary strands having different adaptors) can amplify exponentially.

Having obtained the final differential display, two options are available if cloning of cDNAs is desired. One is to transform the whole of the final PCR reaction into competent cells. Transformed colonies can then be isolated and their inserts characterized by sequencing, restriction analysis or PCR. Alternatively, the final PCR products can be resolved on a gel and the individual bands excised, reamplified and cloned. The first approach is technically simpler and less time consuming. However, ligation/transformation reactions are known to be biased towards the cloning of smaller molecules, and so the final population of clones will probably not contain a representative selection of the larger products. In addition, although equalization theoretically occurs, observations in this laboratory suggest that this is by no means perfectly accomplished. Consequently, some gene species are present in a higher number than others and this will be represented in the final population of clones. Thus, in order to obtain a substantial proportion of those gene species that actually demonstrate differential expression in the tester population, the number of clones that will have to be screened after this step may be substantial. The second approach is initially more time consuming and technically demanding. However, it would appear to offer better prospects for cloning larger and low abundance gel products. In addition, one can incorporate a screening step that differentiates different products of different sequences but of the same size (HA-staining, see later). In this way, a good idea of the final number of clones to be isolated and identified can be achieved.

An alternative (or even complementary) approach is to use the final differential display reaction to screen a cDNA library to isolate full length clones for further characterization, or a DNA array (see later) to quickly identify known genes. SSH has been used in this laboratory to begin characterization of the short-term gene expression profiles of enzyme-inducers such as phenobarbital (Rockett *et al.* 1997) and Wy-14,643 (Rockett *et al.* unpublished observations). The isolation of differentially expressed genes in this manner enables the construction of a fingerprint

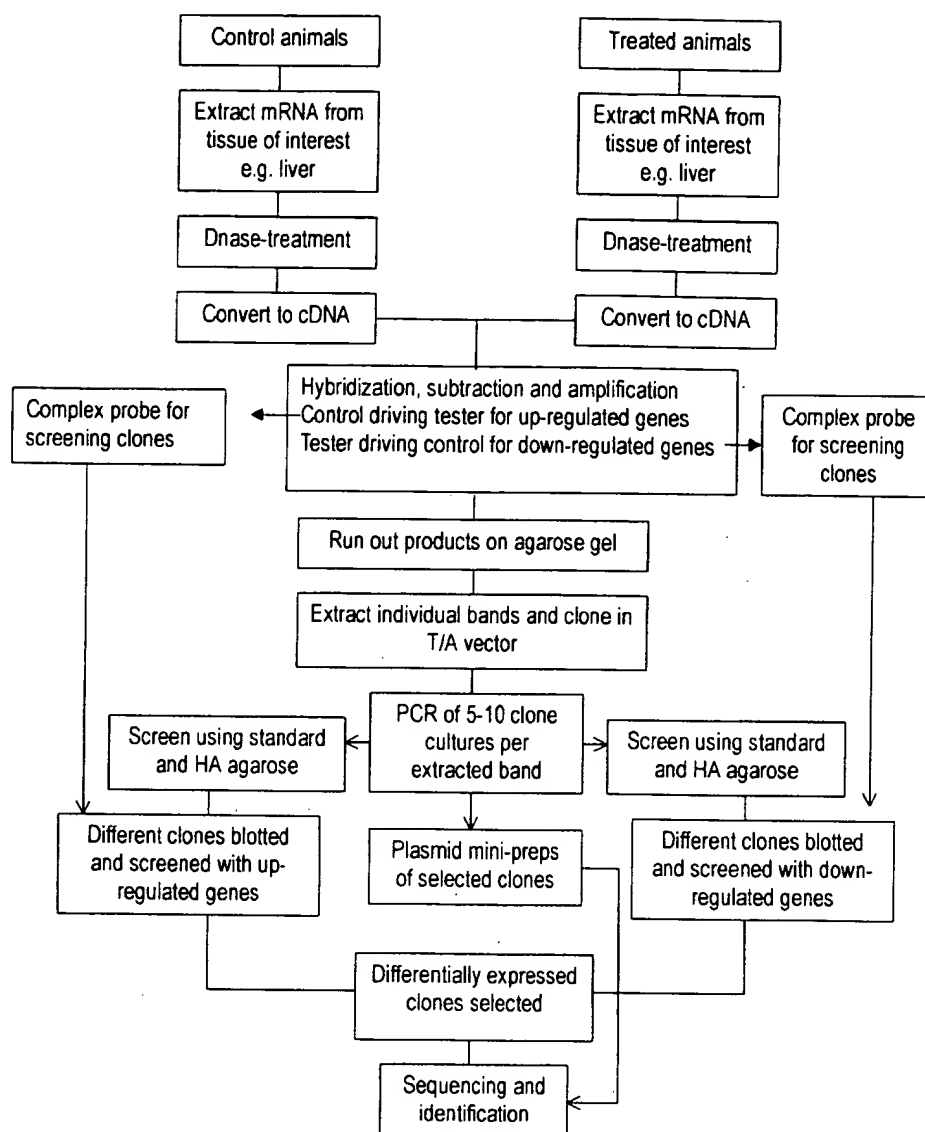


Figure 6. Flow diagram showing method used in this laboratory to isolate and identify clones of genes which are differentially expressed in rat liver following short term exposure to the enzyme inducers, phenobarbital and Wy-14,643.

of expressed genes which are unique to each compound and time/dose point. Such information could be useful in short-term characterization of the toxic potential of new compounds by comparing the gene-expression profiles they elicit with those produced by known inducers. Figure 6 shows a flow diagram of the method used to isolate, verify and clone differentially expressed genes, and figure 7 shows expression profiles obtained from a typical SSH experiment. Subsequent sub-cloning of the individual bands, sequencing and gene data base interrogation reveals many genes which are either up- or down-regulated by phenobarbital in the rat (tables 2 and 3).

One of the advantages in using the SSH approach is that no prior knowledge is required of which specific genes are up/down-regulated subsequent to xenobiotic

Table 2. Genes up-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
5 (1300)	93.5%	CYP2B1
7 (1000)	95.1%	Preproalbumin Serum albumin mRNA
8 (950)	98.3%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
10 (850)	95.7%	CYP2B1
11 (800)	Clone 1 94.9%	CYP2B1
	Clone 2 75.3%	CYP2B2
12 (750)	93.8%	TRPM-2 mRNA
15 (600)	92.9%	Sulfated glycoprotein Preproalbumin Serum albumin mRNA
16 (55)	Clone 1 95.2%	CYP2B1
	Clone 2 93.6%	Haptoglobin mRNA partial alpha
21 (350)	99.3%	18S, 5.8S & 28S rRNA

Bands 1–4, 6, 9, 13, 14, and 17–20 are shown to be false positives by dot blot analysis and, therefore, are not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are up-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

Table 3. Genes down-regulated in rat liver following 3-day exposure to phenobarbital.

Band number (approximate size in bp)	Highest sequence similarity	FASTA-EMBL gene identification
1 (1500)	95.3%	3-oxoacyl-CoA thiolase
2 (1200)	92.3%	Hemopexin mRNA
3 (1000)	91.7%	Alpha-2u-globulin mRNA
7 (700)	Clone 1 77.2%	<i>M. musculus</i> C1 inhibitor
	Clone 2 94.5%	Electron transfer flavoprotein
	Clone 3 91.0%	<i>M. musculus</i> Topoisomerase 1 (Topo 1)
8 (650)	Clone 1 86.9%	Soares 2NbMT <i>M. musculus</i> (EST)
	Clone 2 96.2%	Alpha-2u-globulin (s-type) mRNA
9 (600)	Clone 1 86.9%	Soares mouse NML <i>M. musculus</i> (EST)
	Clone 2 82.0%	Soares p3NMF 19.5 <i>M. musculus</i> (EST)
10 (550)	73.8%	Soares mouse NML <i>M. musculus</i> (EST)
11 (525)	95.7%	NCI-CGAP-Pr1 <i>H. sapiens</i> (EST)
12 (375)	100.0%	Ribosomal protein
13 (23)	Clone 1 97.2%	Soares mouse embryo NbME135 (EST)
	Clone 2 100.0%	Fibrinogen B-beta-chain
	Clone 3 100.0%	Apolipoprotein E gene
14 (170)	96.0%	Soares p3NMF19.5 <i>M. musculus</i> (EST)
15 (140)	97.3%	Stratagene mouse testis (EST)
Others: (300)	96.7%	<i>R. norvegicus</i> RASP 1 mRNA
(275)	93.1%	Soares mouse mammary gland (EST)

EST = Expressed sequence tag. Bands 4–6 were shown to be false positives by dot blot analysis and, therefore, were not sequenced. Derived from Rockett *et al.* (1997). It should be noted that the above genes do not represent the complete spectrum of genes which are down-regulated in rat liver by phenobarbital, but simply represents the genes sequenced and identified to date.

display' (DD). In this method, all the mRNA species in the control and treated cell populations are amplified in separate reactions using reverse transcriptase-PCR (RT-PCR). The products are then run side-by-side on sequencing gels. Those bands which are present in one display only, or which are much more intense in one

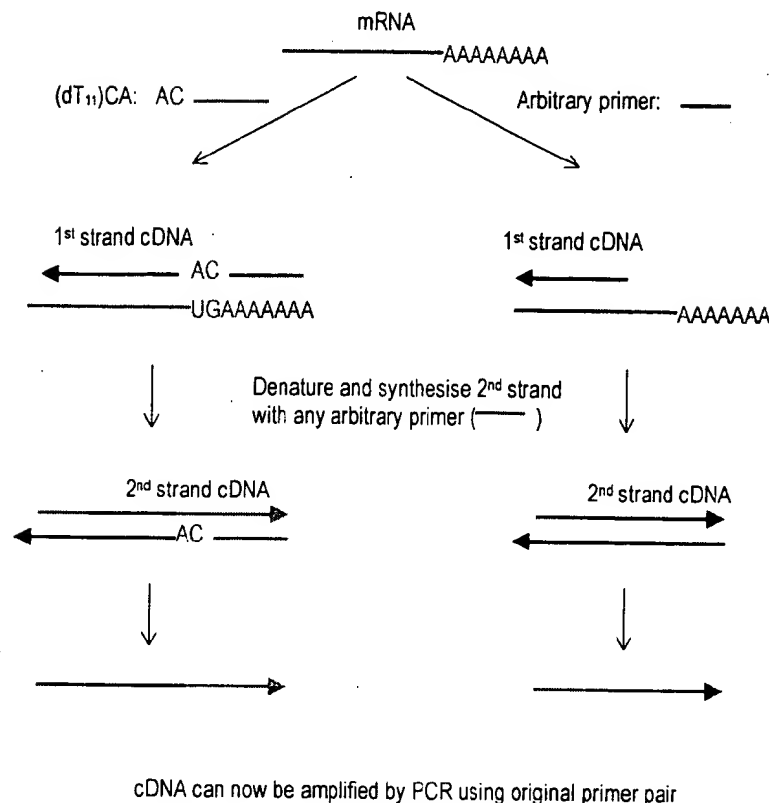


Figure 8. Two approaches to differential display (DD) analysis. 1<sup>st</sup> strand synthesis can be carried out either with a polyd(T)<sub>11</sub>NN primer (where N = G, C or A) or with an arbitrary primer. The use of different combinations of G, C and A to anchor the first strand polyd(T) primer enables the priming of the majority of polyadenylated mRNAs. Arbitrary primers may hybridize at none, one or more places along the length of the mRNA, allowing 1<sup>st</sup> strand cDNA synthesis to occur at none, one or more points in the same gene. In both cases, 2<sup>nd</sup> strand synthesis is carried out with an arbitrary primer. Since these arbitrary primers for the 2<sup>nd</sup> strand may also hybridize to the 1<sup>st</sup> strand cDNA in a number of different places, several different 2<sup>nd</sup> strand products may be obtained from one binding point of the 1<sup>st</sup> strand primer. Following 2<sup>nd</sup> strand synthesis, the original set of primers is used to amplify the second strand products, with the result that numerous gene sequences are amplified.

### Restriction endonuclease-facilitated analysis of gene expression

#### *Serial Analysis of Gene Expression (SAGE)*

A more recent development in the field of differential display is SAGE analysis (Velculescu *et al.* 1995). This method uses a different approach to those discussed so far and is based on two principles. Firstly, in more than 95% of cases, short nucleotide sequences ('tags') of only nine or 10 base pairs provide sufficient information to identify their gene of origin. Secondly, concatenation (linking together in a series) of these tags allows sequencing of multiple cDNAs within a single clone. Figure 9 shows a schematic representation of the SAGE process. In this procedure, double stranded cDNA from the test cells is synthesized with a biotinylated polydT primer. Following digestion with a commonly cutting (4bp recognition sequence) restriction enzyme ('anchoring enzyme'), the 3' ends of the cDNA population are captured with streptavidin beads. The captured population is

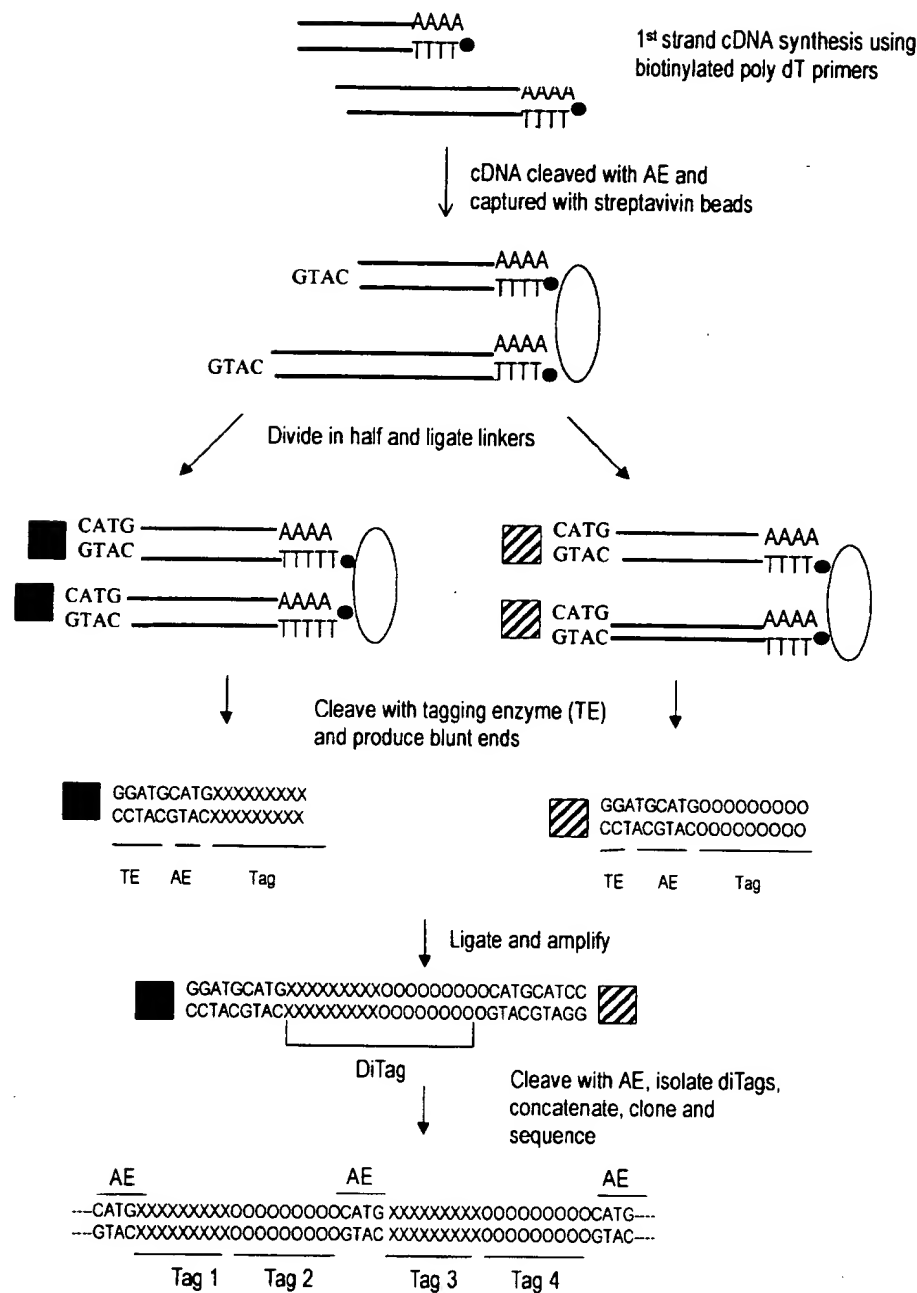


Figure 9. Serial analysis of gene expression (SAGE) analysis. cDNA is cleaved with an anchoring enzyme (AE) and the 3' ends captured using streptavidin beads. The cDNA pool is divided in half and each portion ligated to a different linker, each containing a type IIS restriction site (tagging enzyme, TE). Restriction with the type IIS enzyme releases the linker plus a short length of cDNA (XXXXX and OOOOO indicate nucleotides of different tags). The two pools of tags are then ligated and amplified using linker-specific primers. Following PCR, the products are cleaved with the AE and the ditags isolated from the linkers using PAGE. The ditags are then ligated (during which process, concatenization occurs) and cloned into a vector of choice for sequencing. After Velculescu *et al.* (1995), with permission.

of all human genes (Hillier *et al.* 1996). This large number of freely available sequences (both sequence information and clones are normally available royalty-free from the originators) has enabled the development of a new approach towards differential gene expression analysis as described by Vasmatzis *et al.* (1998). The approach is simple in theory: EST databases are first searched for genes that have a number of related EST sequences from the target tissue of choice, but none or few from non-target tissue libraries. Programmes to assist in the assembly of such sets of overlapping data may be developed in-house or obtained privately or from the internet. For example, the Institute for Genomic Research (TIGR, found at <http://www.tigr.org>) provides many software tools free of charge to the scientific community. Included amongst these is the TIGR assembler (Sutton *et al.* 1995), a tool for the assembly of large sets of overlapping data such as ESTs, bacterial artificial chromosomes (BACs), or small genomes. Candidate EST clones representing different genes are then analysed using RNA blot methods for size and tissue specificity and, if required, used as probes to isolate and identify the full length cDNA clone for further characterization. In practice however, the method is rather more involved, requiring bioinformatic and computer analysis coupled with confirmatory molecular studies. Vasmatzis *et al.* (1998) have described several problems in this fledgling approach, such as separating highly homologous sequences derived from different genes and an overemphasis of specificity for some EST sequences. However, since these problems will largely be addressed by the development of more suitable computer algorithms and an increased completeness of the EST database, it is likely that this approach to identifying differentially expressed genes may enjoy more patronage in the future.

### Problems and potential of differential expression techniques

#### *The holistic or single cell approach?*

When working with *in vivo* models of differential expression, one of the first issues to consider must be the presence of multiple cell types in any given specimen. For example, a liver sample is likely to contain not only hepatocytes, but also (potentially) Ito cells, bile ductule cells, endothelial cells, various immune cells (e.g. lymphocytes, macrophages and Kupffer cells) and fibroblasts. Other tissues will each have their own distinctive cell populations. Also, in the case of neoplastic tissue, there are almost always normal, hyperplastic and/or dysplastic cells present in a sample. One must, therefore, be aware that genes obtained from a differential display experiment performed on an animal tissue model may not necessarily arise exclusively from the intended 'target' cells, e.g. hepatocytes/neoplastic cells. If appropriate, further analyses using immunohistochemistry, *in situ* hybridization or *in situ* RT-PCR should be used to confirm which cell types are expressing the gene(s) of interest. This problem is probably most acute for those studying the differential expression of genes in the development of different cell types, where there is a need to examine homologous cell populations. The problem is now being addressed at the National Cancer Institute (Bethesda, MD, USA) where new micro-dissection techniques have been employed to assist in their gene analysis programme, the Cancer Genome Anatomy Project (CGAP) (For more information see web site: <http://www.ncbi.nlm.nih.gov/ncicgap/intro.html>). There are also separation techniques available that utilise cell-specific antigens as a means to isolate target cells,



species present at less than 1.2% of the total mRNA population—equivalent to an intermediate or abundant species. Interestingly, when simple model systems (single target only) were used instead of a heterogeneous mRNA population, the same primers could detect levels of target mRNA down to 10000× smaller. These results are probably best explained by competition for substrates from the many PCR products produced in a DD reaction.

The numbers of differentially expressed mRNAs reported in the literature using various model systems provides further evidence that many differentially expressed mRNAs are not recovered. For example, DeRisi *et al.* (1997) used DNA array technology to examine gene expression in yeast following exhaustion of sugar in the medium, and found that more than 1700 genes showed a change in expression of at least 2-fold. In light of such a finding, it would not be unreasonable to suggest that of the 8000–15 000 different mRNA species produced by any given mammalian cell, up to 1000 or more may show altered expression following chemical stimulation. Whilst this may be an extreme figure, it is known that at least 100 genes are activated/upregulated in Jurkat (T-) cells following IL-2 stimulation (Ullman *et al.* 1990). In addition, Wan *et al.* (1996) estimated that interferon- $\gamma$ -stimulated HeLa cells differentially express up to 433 genes (assuming 24000 distinct mRNAs expressed by the cells). However, there have been few publications documenting anywhere near the recovery of these numbers. For example, in using DD to compare normal and regenerating mouse liver, Bauer *et al.* (1993) found only 70 of 38000 total bands to be different. Of these, 50% (35 genes) were shown to correspond to differentially expressed bands. Chen *et al.* (1996) reported 10 genes upregulated in female rat liver following ethinyl estradiol treatment. McKenzie and Drake (1997) identified 14 different gene products whose expression was altered by phorbol myristate acetate (PMA, a tumour promoter agent) stimulation of a human myelomonocytic cell line. Kilty and Vickers (1997) identified 10 different gene products whose expression was upregulated in the peripheral blood leukocytes of allergic disease sufferers. Linskens *et al.* (1995) found 23 genes differentially expressed between young and senescent fibroblasts. Techniques other than DD have also provided an apparent paucity of differentially expressed genes. Using SH for example, Cao *et al.* (1997) found 15 genes differentially expressed in colorectal cancer compared to normal mucosal epithelium. Fitzpatrick *et al.* (1995) isolated 17 genes upregulated in rat liver following treatment with the peroxisome proliferator, clofibrate; Philips *et al.* (1990) isolated 12 cDNA clones which were upregulated in highly metastatic mammary adenocarcinoma cell lines compared to poorly metastatic ones. Prashar and Weissman (1996) used 3' restriction fragment analysis and identified approximately 40 genes showing altered expression within 4 h of activation of Jurkat T-cells. Groenink and Leegwater (1996) analysed 27 gene fragments isolated using SSH of delayed early response phase of liver regeneration and found only 12 to be upregulated.

In the laboratory, SSH was used to isolate up to 70 candidate genes which appear to show altered expression in guinea pig liver following short-term treatment with the peroxisome proliferator, WY-14,643 (Rockett, Swales, Esdaile and Gibson, unpublished observations). However, these findings have still to be confirmed by analysis of the extracted tissue mRNA for differential expression of these sequences.

Whilst the latest differential display technologies are purported to include design and experimental modifications to overcome this lack of efficiency (in both the total number of differentially expressed genes recovered and the percentage that are true

experiments and animals. DD, on the other hand, is not subject to this grey zone since, unlike SH approaches, it does not amplify the difference in expression between two samples. Wan *et al.* (1996) reported that differences in expression of twofold or more are detectable using DD.

#### *Resolution and visualization of differential expression products*

It seems highly improbable with current technology that a gel system could be developed that is able to resolve all gene species showing altered expression in any given test system (be it SH- or DD-based). Polyacrylamide gel electrophoresis (PAGE) can resolve size differences down to 0.2% (Sambrook *et al.* 1989) and are used as standard in DD experiments. Even so, it is clear that a complex series of gene products such as those seen in a DD will contain unresolvable components. Thus, what appears to be one band in a gel may in fact turn out to be several. Indeed, it has been well documented (Mathieu-Daude *et al.* 1996, Smith *et al.* 1997) that a single band extracted from a DD often represents a composite of heterogeneous products, and the same has been found for SSH displays in this laboratory (Rockett *et al.* 1997). One possible solution was offered by Mathieu-Daude *et al.* (1996), who extracted and reamplified candidate bands from a DD display and used single strand conformation polymorphism (SSCP) analysis to confirm which components represented the truly differentially expressed product.

Many scientists often try to avoid the use of PAGE where possible because it is technically more demanding than agarose gel electrophoresis (AGE). Unfortunately, high resolution agarose gels such as Metaphor (FMC, Lichfield, UK) and AquaPor HR (National Diagnostics, Hesse, UK), whilst easier to prepare and manipulate than PAGE, can only separate DNA sequences which differ in size by around 1.5–2% (15–20 base pairs for a 1Kb fragment). Thus, SSH, RDA or other such products which differ in size by less than this amount are normally not resolvable. However, a simple technique does in fact exist for increasing the resolving power of AGE—the inclusion of HA-red (10-phenyl neutral red-PEG ligand) or HA-yellow (bisbenzamide-PEG ligand) (Hanse Analytik GmbH, Bremen, Germany) in a gel separates identical or closely sized products on base content. Specifically, HA-red and -yellow selectively bind to GC and AT DNA motifs, respectively (Wawer *et al.* 1995, Hanse Analytik 1997, personal communication). Since both HA-stains possess an overall positive charge, they migrate towards the cathode when an electric field is applied. This is in direct opposition to DNA, which is negatively charged and, therefore, migrates towards the anode. Thus, if two DNA clones are identical in size (as perceived on a standard high resolution agarose gel), but differ in AT/GC content, inclusion of a HA-dye in the gel will effectively retard the migration of one of the sequences compared to the other, effectively making it apparently larger and, thus, providing a means of differentiating between the two. The use of HA-red has been shown to resolve sequences with an AT variation of less than 1% (Wawer *et al.* 1995), whilst Hanse Analytik have reported that HA staining is so sensitive that in one case it was used to distinguish two 567bp sequences which differed by only a single point mutation (Hanse Analytik 1996, personal communication). Therefore, if one wishes to check whether all the clones produced from a specific band in a differential display experiment are derived from the same gene species, a small amount of reamplified or digested clone can be run on a standard high resolution gel, and a second aliquot

Extraction of differentially expressed bands from a gel can be complex since, in some cases (e.g. DD, GEF), the results are visualized by autoradiographic means, such that precise overlay of the developed film on the gel must occur if the correct band is to be extracted for further analysis. Clearly, a misjudged extraction can account for many man-hours lost. This problem, and that of the use of radioisotopes, has been addressed by several groups. For example, Lohmann *et al.* (1995) demonstrated that silver staining can be used directly to visualize DD bands in horizontal PAGs. An *et al.* (1996) avoided the use of radioisotopes by transferring a small amount (20–30%) of the DNA from their DD to a nylon membrane, and visualizing the bands using chemiluminescent staining before going back to extract the remaining DNA from the gel. Chen and Peck (1996) went one step further and transferred the entire DD to a nylon membrane. The DNA bands were then visualized using a digoxigenin (DIG) system (DIG was attached to the polydT primers used in the differential display procedure). Differentially expressed bands were cut from the membrane and the DNA eluted by washing with PCR buffer prior to reamplification.

One of the advantages of using techniques such as SSH and RDA is that the final display can be run on an agarose gel and the bands visualized with simple ethidium bromide staining. Whilst this approach can provide acceptable results, overstaining with SYBR Green I or SYBR Gold nucleic acid stains (FMC) effectively enhances the intensity and sharpness of the bands. This greatly aids in their precise extraction and often reveals some faint products that may otherwise be overlooked. Whilst differential displays stained with SYBR Green I are better visualized using short wavelength UV (254 nm) rather than medium wavelength (306 nm), the shorter wavelength is much more DNA damaging. In practice, it takes only a few seconds to damage DNA extracted under 254 nm irradiation, effectively preventing reamplification and cloning. The best approach is to overstain with SYBR Green I and extract bands under a medium wavelength UV transillumination.

#### **The possible use of 'microfingerprinting' to reduce complexity**

Given the sheer number of gene products and the possible complexity of each band, an alternative approach to rapid characterization may be to use an enhanced analysis of a small section of a differential display—a 'sub-fingerprint' or 'micro-fingerprint'. In this case, one could concentrate on those bands which only appear in a particular chosen size region. Reducing the fingerprint in this way has at least two advantages. One is that it should be possible to use different gel types, concentrations and run times tailored exactly to that region. Currently, one might run products from 100–3000 + bp on the same gel, which leads to compromise in the gel system being used and consequently to suboptimal resolution, both in terms of size and numbers, and can lead to problems in the accurate excision of individual bands. Secondly, it may be possible to enhance resolution by using a 2-D analysis using a HA-stain, as described earlier. In summary, if a range of gene product sizes is carefully chosen to include certain 'relevant' genes, the 2-D system standardized, and appropriate gene analysis used, it may be possible to develop a method for the early and rapid identification of compounds which have similar or widely different cellular effects. If the prognosis for exposure to one or more other chemicals which display a similar profile is already known, then one could perhaps predict similar effects for any new compounds which show a similar micro-fingerprint.

which was up-regulated in the liver of rats exposed to Wy-14,643 and was identified by a FASTA search as being transferrin (data not shown). However, transferrin is known to be downregulated by hypolipidemic peroxisome proliferators such as Wy-14,643 (Hertz *et al.* 1996), and this was confirmed with subsequent RT-PCR analysis. This suggests that the gene sequence isolated may belong to a gene which is closely related to transferrin, but is regulated by a different mechanism.

A further problem associated with SH technology is redundancy. In most cases before SH is carried out, the cDNA population must first be simplified by restriction digestion. This is important for at least two reasons:

- (1) To reduce complexity—long cDNA fragments may form complex networks which prevent the formation of appropriate hybrids, especially at the high concentrations required for efficient hybridization.
- (2) Cutting the cDNAs into small fragments provides better representation of individual genes. This is because genes derived from related but distinct members of gene families often have similar coding sequences that may cross-hybridize and be eliminated during the subtraction procedure (Ko 1990). Furthermore, different fragments from the same cDNA may differ considerably in terms of hybridization and amplification and, thus, may not efficiently do one or the other (Wang and Brown 1991). Thus, some fragments from differentially expressed cDNAs may be eliminated during subtractive hybridization procedures. However, other fragments may be enriched and isolated. As a consequence of this, some genes will be cut one or more times, giving rise to two or more fragments of different sizes. If those same genes are differentially expressed, then two or more of the different size fragments may come through as separate bands on the final differential display, increasing the observed redundancy and increasing the number of redundant sequencing reactions.

Sequence comparisons also throw up another important point—at what degree of sequence similarity does one accept a result. Is 90% identity between a gene derived from your model species and another acceptably close? Is 95% between your sequence and one from the same species also acceptable? This problem is particularly relevant when the forward and reverse sequence comparisons give similar sequences with completely different gene species! An arbitrary decision seems to be to allocate genes that are definite (95% and above similarity) and then group those between 60 and 95% as being related or possible homologues.

### Quantitative analysis

At some point, one must give consideration to the quantitative analysis of the candidate genes, either as a means of confirming that they are truly differentially expressed, or in order to establish just what the differences are. Northern blot analysis is a popular approach as it is relatively easy and quick to perform. However, the major drawback with Northern blots is that they are often not sensitive enough to detect rare sequences. Since the majority of messages expressed in a cell are of low abundance (see table 1), this is a major problem. Consequently, RT-PCR may be the method of choice for confirming differential expression. Although the procedure is somewhat more complex than Northern analysis, requiring synthesis of primers and optimization of reaction conditions for each gene species, it is now possible to set up high throughput PCR systems using multichannel pipettes, 96 +-well plates and

become clear that the fingerprinting process, whilst still valid, is much too complex to be represented by a single technique profile. This is because all differential display techniques have common and/or unique technical problems which preclude the isolation and identification of all those genes which show changes in expression. Furthermore, there are important genetic changes related to disease development which differential expression analysis is simply not designed to address. An example of this is the presence of small deletions, insertions, or point mutations such as those seen in activated oncogenes, tumour suppressor genes and individual polymorphisms. Polymorphic variations, small though they usually are, are often regarded as being of paramount importance in explaining why some patients respond better than others to certain drug treatments (and, in logical extension, why some people are less affected by potentially dangerous xenobiotics/carcinogens than others). The identification of such point mutations and naturally occurring polymorphisms requires the subsequent application of sequencing, SSCP, DGGE or TGGE to the gene of interest. Furthermore, differential display is not designed to address issues such as alternatively spliced gene species or whether an increased abundance of mRNA is a result of increased transcription or increased mRNA stability.

### *Conclusions*

Perhaps the main advantage of open system differential display techniques is that they are not limited by extant theories or researcher bias in revealing genes which are differentially expressed, since they are designed to amplify all genes which demonstrate altered expression. This means that they are useful for the isolation of previously unknown genes which may turn out be useful biomarkers of a particular state or condition. At least one open system (SAGE) is also quantitative, thus eliminating the need to return to the original mRNA and carry out Northern/PCR analysis to confirm the result. However, the rapid progress of genome mapping projects means that over the next 5–10 years or so, the balance of experimental use will switch from open to closed differential display systems, particularly DNA arrays. Arrays are easier and faster to prepare and use, provide quantitative data, are suitable for high throughput analysis and can be tailored to look at specific signalling pathways or families of genes. Identification of all the gene sequences in human and common laboratory animals combined with improved DNA array technology, means that it will soon no longer be necessary to try to isolate differentially expressed genes using the technically more demanding open system approach. Thus, their main advantage (that of identifying unknown genes) will be largely eradicated. It is likely, therefore, that their sphere of application will be reduced to analysis of the less common laboratory species, since it will be some time yet before the genomes of such animals as zebrafish, electric eels, gerbils, crayfish and squid, for example, will be sequenced.

Of course, in the end the question will always remain: What is the functional/biological significance of the identified, differentially expressed genes? One persistent problem is understanding whether differentially expressed genes are a cause or consequence of the altered state. Furthermore, many chemicals, such as non-genotoxic carcinogens, are also mitogens and so genes associated with replication will also be upregulated but may have little or nothing to do with the

US Environmental Protection Agency and approved for publication. Approval does not signify that the contents reflect the views and policies of the Agency, nor does mention of trade names constitute endorsement or recommendation for use.

## References

- ADAMS, M. D., KELLEY, J. M., GOCAYNE, J. D., DUBNICK, M., POLYMERPOULOS, M. H., XIAO, H., MERRIL, C. R., WU, A., OLDE, B., MORENO, R. F., KERLAVAGE, A. R., MCCOMBIE, W. R. and VENTOR, J. C., 1991, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651-1656.
- AN, G., LUO, G., VELTRI, R. W. and O'HARA, S. M., 1996, Sensitive non-radioactive differential display method using chemiluminescent detection. *Biotechniques*, **20**, 342-346.
- AXEL, R., FEIGELSON, P. and SCHULTZ, G., 1976, Analysis of the complexity and diversity of mRNA from chicken liver and oviduct. *Cell*, **7**, 247-254.
- BAND, V. and SAGER, R., 1989, Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proceedings of the National Academy of Sciences, USA*, **86**, 1249-1253.
- BAUER, D., MULLER, H., REICH, J., RIEDEL, H., AHRENKIEL, V., WARTHOF, P. and STRAUSS, M., 1993, Identification of differentially expressed mRNA species by an improved display technique (DDRT-PCR). *Nucleic Acids Research*, **21**, 4272-4280.
- BERTIOLI, D. J., SCHLICHTER, U. H. A., ADAMS, M. J., BURROWS, P. R., STEINBISS, H.-H. and ANTONIW, J. F., 1995, An analysis of differential display shows a strong bias towards high copy number mRNAs. *Nucleic Acids Research*, **23**, 4520-4523.
- BRAVO, R., 1990, Genes induced during the G0/G1 transition in mouse fibroblasts. *Seminars in Cancer Biology*, **1**, 37-46.
- BURN, T. C., PETROVICK, M. S., HOHAUS, S., ROLLINS, B. J. and TENEN, D. G., 1994, Monocyte chemoattractant protein-1 gene is expressed in activated neutrophils and retinoic acid-induced human myeloid cell lines. *Blood*, **84**, 2776-2783.
- CAO, J., CAI, X., ZHENG, L., GENG, L., SHI, Z., PAO, C. C. and ZHENG, S., 1997, Characterisation of colorectal cancer-related cDNA clones obtained by subtractive hybridisation screening. *Journal of Cancer Research and Clinical Oncology*, **123**, 447-451.
- CASSIDY, S. B., 1995, Uniparental disomy and genomic imprinting as causes of human genetic disease. *Environmental and Molecular Mutagenesis*, **25** (Suppl 26), 13-20.
- CHANG, G. W. and TERZAGHI-HOWE, M., 1998, Multiple changes in gene expression are associated with normal cell-induced modulation of the neoplastic phenotype. *Cancer Research*, **58**, 4445-4452.
- CHEN, J., SCHWARTZ, D. A., YOUNG, T. A., NORRIS, J. S. and YAGER, J. D., 1996, Identification of genes whose expression is altered during mitosuppression in livers of ethinyl estradiol-treated female rats. *Carcinogenesis*, **17**, 2783-2786.
- CHEN, J. J. W. and PECK, K., 1996, Non-radioactive differential display method to directly visualise and amplify differential bands on nylon membrane. *Nucleic Acid Research*, **24**, 793-794.
- CLON TECHNIQUES, 1997a, PCR-Select Differential Screening Kit—the next step after Clontech PCR-Select cDNA subtraction. *ClonTechniques*, **XII**, 18-19.
- CLON TECHNIQUES, 1997b, Housekeeping RT-PCR amplimers and cDNA probes. *ClonTechniques*, **XII**, 15-16.
- DAVIS, M. M., COHEN, D. I., NIELSEN, E. A., STEINMETZ, M., PAUL, W. E. and HOOD, L., 1984, Cell-type-specific cDNA probes and the murine I region: the localization and orientation of Ad alpha. *Proceedings of the National Academy of Sciences (USA)*, **81**, 2194-2198.
- DELLAVALLE, R. P., PETERSON, R. and LINDQUIST, S., 1994, Preferential deadenylation of HSP70 mRNA plays a key role in regulating Hsp70 expression in *Drosophila melanogaster*. *Molecular and Cell Biology*, **14**, 3646-3659.
- DERISI, J. L., VASHWANATH, R. L. and BROWN, P., 1997, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680-686.
- DIATCHENKO, L., LAU, Y.-F. C., CAMPBELL, A. P., CHENCHIK, A., MOQADAM, F., HUANG, B., LUKYANOV, K., GURSKAYA, N., SVERDLOV, E. D. and SIEBERT, P. D., 1996, Suppression subtractive hybridisation: A method for generating differentially regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences (USA)*, **93**, 6025-6030.
- DOGRA, S. C., WHITELAW, M. L. and MAY, B. K., 1998, Transcriptional activation of cytochrome P450 genes by different classes of chemical inducers. *Clinical and Experimental Pharmacology and Physiology*, **25**, 1-9.
- DUGUID, J. R. and DINAUER, M. C., 1990, Library subtraction of *in vitro* cDNA libraries to identify differentially expressed genes in scrapie infection. *Nucleic Acids Research*, **18**, 2789-2792.
- DUNBAR, P. R., OGG, G. S., CHEN, J., RUST, N., VAN DER BRUGGEN, P. and CERUNDOLO, V., 1998, Direct isolation, phenotyping and cloning of low-frequency antigen-specific cytotoxic T lymphocytes from peripheral blood. *Current Biology*, **26**, 413-416.

- quantifying cytomegalic endothelial cells in peripheral blood from cytomegalovirus-infected patients. *Clinical Diagnostic and Laboratory Immunology*, **5**, 622-626.
- KILTY, I. and VICKERS, P., 1997, Fractionating DNA fragments generated by differential display PCR. *Strategies Newsletter* (Stratagene), **10**, 50-51.
- KLEINJAN, D.-J. and VAN HEYNINGEN, V., 1998, Position effect in human genetic disease. *Human and Molecular Genetics*, **7**, 1611-1618.
- KO, M. S., 1990, An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research*, **18**, 5705-5711.
- LAKE, B. G., EVANS, J. G., CUNNINGHAM, M. E. and PRICE, R. J., 1993, Comparison of the hepatic effects of Wy-14,643 on peroxisome proliferation and cell replication in the rat and Syrian hamster. *Environmental Health Perspectives*, **101**, 241-248.
- LAKE, B. G., EVANS, J. G., GRAY, T. J. B., KOROSI, S. A. and NORTH, C. J., 1989, Comparative studies of nafenopin-induced hepatic peroxisome proliferation in the rat, Syrian hamster, guinea pig and marmoset. *Toxicology and Applied Pharmacology*, **99**, 148-160.
- LENNARD, M. S., 1993, Genetically determined adverse drug reactions involving metabolism. *Drug Safety*, **9**, 60-77.
- LEVY, S., TODD, S. C. and MAECKER, H. T., 1998, CD81(TAPA-1): a molecule involved in signal transduction and cell adhesion in the immune system. *Annual Review of Immunology*, **16**, 89-109.
- LIANG, P. and PARDEE, A. B., 1992, Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967-971.
- LIANG, P., AVERBOUKH, L., KEYOMARSI, K., SAGER, R. and PARDEE, A., 1992, Differential display and cloning of messenger RNAs from human breast cancer versus mammary epithelial cells. *Cancer Research*, **52**, 6966-6968.
- LIANG, P., AVERBOUKH, L. and PARDEE, A. B., 1993, Distribution & cloning of eukaryotic mRNAs by means of differential display refinements and optimisation. *Nucleic Acids Research*, **21**, 3269-3275.
- LIANG, P., BAUER, D., AVERBOUKH, L., WARTHOF, P., ROHRWILD, M., MULLER, H., STRAUSS, M. and PARDEE, A. B., 1995, Analysis of altered gene expression by differential display. *Methods in Enzymology*, **254**, 304-321.
- LINSKENS, M. H., FENG, J., ANDREWS, W. H., ENLOW, B. E., SAATI, S. M., TONKIN, L. A., FUNK, W. D. and VILLEPONTEAU, B., 1995, Cataloging altered gene expression in young and senescent cells using enhanced differential display. *Nucleic Acids Research*, **23**, 3244-3251.
- LISITSYN, N., LISITSYN, N. and WIGLER, M., 1993, Cloning the differences between two complex genomes. *Science*, **259**, 946-951.
- LOHMANN, J., SCHICKLE, H. and BOSCH, T. C. G., 1995, REN Display, a rapid and efficient method for non-radioactive differential display and mRNA isolation. *Biotechniques*, **18**, 200-202.
- LUNNEY, J. K., 1998, Cytokines orchestrating the immune response. *Reviews in Science and Technology*, **17**, 84-94.
- MAKOWSKA, J. M., GIBSON, G. G. and BONNER, F. W., 1992, Species differences in ciprofibrate-induction of hepatic cytochrome P450A1 and peroxisome proliferation. *Journal of Biochemical Toxicology*, **7**, 183-191.
- MALDARELLI, F., XIANG, C., CHAMOUN, G. and ZEICHNER, S. L., 1998, The expression of the essential nuclear splicing factor SC35 is altered by human immunodeficiency virus infection. *Virus Research*, **53**, 39-51.
- MATHIEU-DAUDE, F., CHENG, R., WELSH, J. and MCCLELLAND, M., 1996, Screening of differentially amplified cDNA products from RNA arbitrarily primed PCR fingerprints using single strand conformation polymorphism (SSCP) gels. *Nucleic Acids Research*, **24**, 1504-1507.
- MCKENZIE, D. and DRAKE, D., 1997, Identification of differentially expressed gene products with the castaway system. *Strategies Newsletter* (Stratagene), **10**, 19-20.
- MCCLELLAND, M., MATHIEU-DAUDE, F. and WELSH, J., 1996, RNA fingerprinting and differential display using arbitrarily primed PCR. *Trends in Genetics*, **11**, 242-246.
- MECHLER, B. and RABBITTS, T. H., 1981, Membrane-bound ribosomes of myeloma cells. IV. mRNA complexity of free and membrane-bound polysomes. *Journal of Cell Biology*, **88**, 29-36.
- MEYER, U. A. and ZANGER, U. M., 1997, Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annual Review of Pharmacology and Toxicology*, **37**, 269-296.
- MOHLER, K. M. and BUTLER, L. D., 1991, Quantitation of cytokine mRNA levels utilizing the reverse transcriptase-polymerase chain reaction following primary antigen-specific sensitization in vivo—I. Verification of linearity, reproducibility and specificity. *Molecular Immunology*, **28**, 437-447.
- MURPHY, L. D., HERZOG, C. E., RUDICK, J. B., TITO FOJO, A. and BATES, S. E., 1990, Use of the polymerase chain reaction in the quantitation of the *mdr-1* gene expression. *Biochemistry*, **29**, 10351-10356.
- NELSON, D. R., KOYMANS, L., KAMATAKI, T., STEGEMAN, J. J., FEYEREISEN, R., WAXMAN, D. J., WATERMAN, M. R., GOTOH, O., COON, M. J., ESTABROOK, R. W., GUNSALUS, I. C. and NEBERT, D. W., 1996, Update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics*, **6**, 1-42.

- SUNG, Y. J. and DENMAN, R. B., 1997, Use of two reverse transcriptases eliminates false-positive results in differential display. *Biotechniques*, **23**, 462-464.
- SUTTON, G., WHITE, O., ADAMS, M. and KERLAVAGE, A., 1995, TIGR Assembler; A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, **1**, 9-19.
- SUZUKI, Y., SEKIYA, T. and HAYASHI, K., 1991, Allele-specific polymerase chain reaction: a method for amplification and sequence determination of a single component among a mixture of sequence variants. *Analytical Biochemistry*, **192**, 82-84.
- SYED, V., GU, W. and HECHT, N. B., 1997, Sertoli cells in culture and mRNA differential display provide a sensitive early warning assay system to detect changes induced by xenobiotics. *Journal of Andrology*, **18**, 264-273.
- UITTERLINDEN, A. G., SLAGBOOM, P., KNOOK, D. L. and VIJGL, J., 1989, Two-dimensional DNA fingerprinting of human individuals. *Proceedings of the National Academy of Sciences (USA)*, **86**, 2742-2746.
- ULLMAN, K. S., NORTHROP, J. P., VERWEIJ, C. L. and CRABTREE, G. R., 1990, Transmission of signals from the T lymphocyte antigen receptor to the genes responsible for cell proliferation and immune function: the missing link. *Annual Review of Immunology*, **8**, 421-452.
- VASMATZIS, G., ESSAND, M., BRINKMANN, U., LEE, B. and PASTON, I., 1998, Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proceedings of the National Academy of Sciences (USA)*, **95**, 300-304.
- VELCULESCU, V. E., ZHANG, L., VOGELSTEIN, B. and KINZLER, K. W., 1995, Serial analysis of gene expression. *Science*, **270**, 484-487.
- VOELTZ, G. K. and STEITZ, J. A., 1998, AuuuA sequences direct mRNA deadenylation uncoupled from decay during *Xenopus* early development. *Molecular and Cell Biology*, **18**, 7537-7545.
- VOGELSTEIN, B. and KINZLER, K. W., 1993, The multistep nature of cancer. *Trends in Genetics*, **9**, 138-141.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WAN, J. S., SHARP, S. J., POIRIER, G. M.-C., WAGAMAN, P. C., CHAMBERS, J., PYATI, J., HOM, Y.-L., GALINDO, J. E., HUVAR, A., PETERSON, P. A., JACKSON, M. R. and ERLANDER, M. G., 1996, Cloning differentially expressed mRNAs. *Nature Biotechnology*, **14**, 1685-1691.
- WALTER, J., BELFIELD, M., HAMPSON, I. and READ, C., 1997, A novel approach for generating subtractive probes for differential screening by CCLS. *Life Science News*, **21**, 13-14.
- WANG, Z. and BROWN, D. D., 1991, A gene expression screen. *Proceedings of the National Academy of Sciences (USA)*, **88**, 11505-11509.
- WAWER, C., RUGGEBERG, H., MEYER, G. and MUYZER, G., 1995, A simple and rapid electrophoresis method to detect sequence variation in PCR-amplified DNA fragments. *Nucleic Acids Research*, **23**, 4928-4929.
- WELSH, J., CHADA, K., DALAL, S. S., CHENG, R., RALPH, D. and MCCLELLAND, M., 1992, Arbitrarily primed PCR fingerprinting of RNA. *Nucleic Acids Research*, **20**, 4965-4970.
- WONG, H., ANDERSON, W. D., CHENG, T. and RIABOWOL, K. T., 1994, Monitoring mRNA expression by polymerase chain reaction: the 'primer-dropping' method. *Analytical Biochemistry*, **223**, 251-258.
- WONG, K. K. and MCCLELLAND, M., 1994, Stress-inducible gene of *Salmonella typhimurium* identified by arbitrarily primed PCR of RNA. *Proceedings of the National Academy of Sciences (USA)*, **91**, 639-643.
- WYNFORD-THOMAS, D., 1991, Oncogenes and anti-oncogenes; the molecular basis of tumour behaviour. *Journal of Pathology*, **165**, 187-201.
- XHU, D., CHAN, W. L., LEUNG, B. P., HUANG, F. P., WHEELER, R., PIEDRAFITA, D., ROBINSON, J. H. and LIEW, F. Y., 1998, Selective expression of a stable cell surface molecule on type 2 but not type 1 helper T cells. *Journal of Experimental Medicine*, **187**, 787-794.
- YANG, M. and SYTOWSKI, A. J., 1996, Cloning differentially expressed genes by linker capture subtraction. *Analytical Biochemistry*, **237**, 109-114.
- ZHAO, N., HASHIDA, H., TAKAHASHI, N., MISUMI, Y. and SAKAKI, Y., 1995, High-density cDNA filter analysis: a novel approach for large scale quantitative analysis of gene expression. *Gene*, **156**, 207-213.
- ZHAO, X. J., NEWSOME, J. T. and CIHLAR, R. L., 1998, Up-regulation of two *Candida albicans* genes in the rat model of oral candidiasis detected by differential display. *Microbial Pathogenesis*, **25**, 121-129.
- ZIMMERMANN, C. R., ORR, W. C., LECLERC, R. F., BARNARD, C. and TIMBERLAKE, W. E., 1980, Molecular cloning and selection of genes regulated in *Aspergillus* development. *Cell*, **21**, 709-715.



## Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI\*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS\*§

\*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

**ABSTRACT** The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely

important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5' end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1997 by The National Academy of Sciences 0027-8424/97/948945-3\$2.00/0  
PNAS is available online at <http://www.pnas.org>.

†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: [gilbert@cmsgm.stanford.edu](mailto:gilbert@cmsgm.stanford.edu).

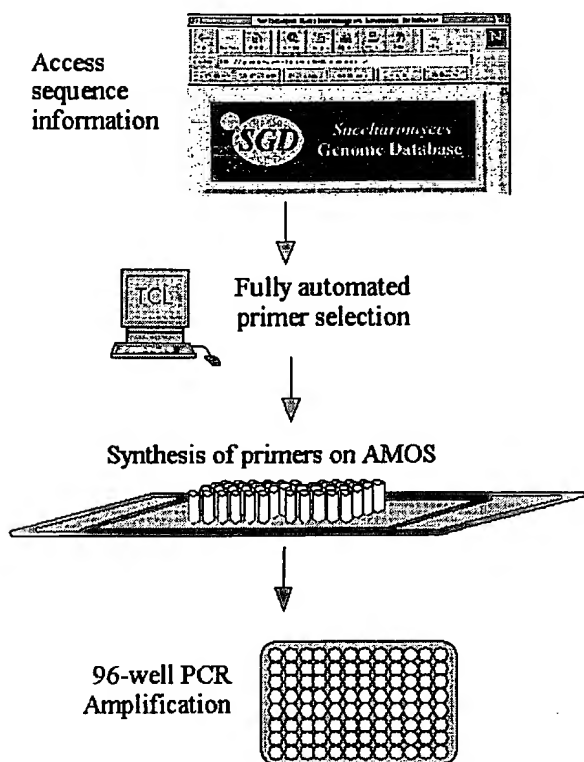


FIG. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose.

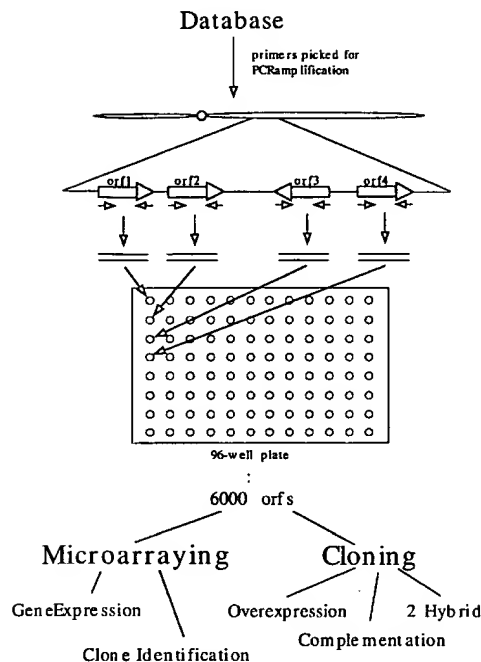


FIG. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a “snapshot” of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

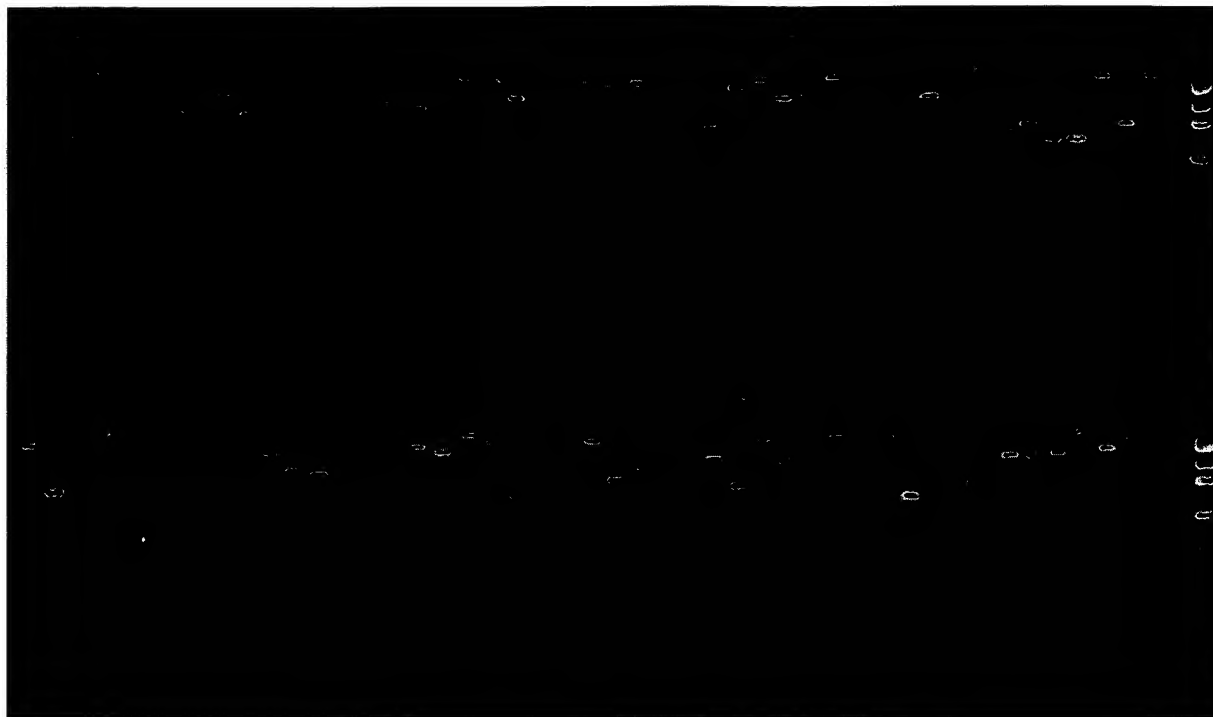


FIG. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct read-out. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

Support was provided by National Institutes of Health Grants R37H60198 and P01H600205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269**, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270**, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273**, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356**, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106**, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379**, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57**, A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251**, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274**, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14**, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274**, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316**, 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* **340**, 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* **4**, 11–18.

## Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,<sup>1</sup> Michael Bittner,<sup>2</sup> Jeffrey Trent,<sup>2</sup> J. Carl Barrett,<sup>1</sup> and Cynthia A. Afshari<sup>1</sup>

<sup>1</sup>Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

<sup>2</sup>Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog.* 24:153-159, 1999. © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

### INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae*, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cerevisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNase protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10-12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

### MICROARRAY DEVELOPMENT AND APPLICATIONS

#### cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

\*Correspondence to: Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, 111 Alexander Drive, Research Triangle Park, NC 27709.

Received 8 December 1998; Accepted 5 January 1999

Abbreviations: PAH, polycyclic aromatic hydrocarbon; NIEHS, National Institute of Environmental Health Sciences.

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluorophores. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease-related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cerevisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

#### Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only  $4n$  cycles (where  $n$  = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)<sup>+</sup> RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and *BRCA1* [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

#### THE USE OF MICROARRAYS IN TOXICOLOGY

##### Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.

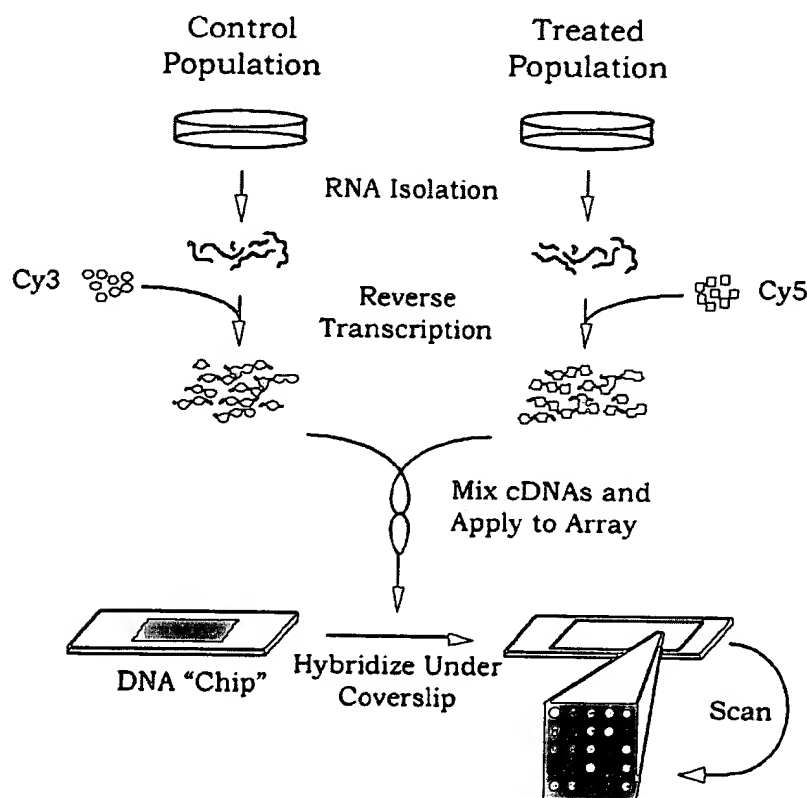


Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illus-

trative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.

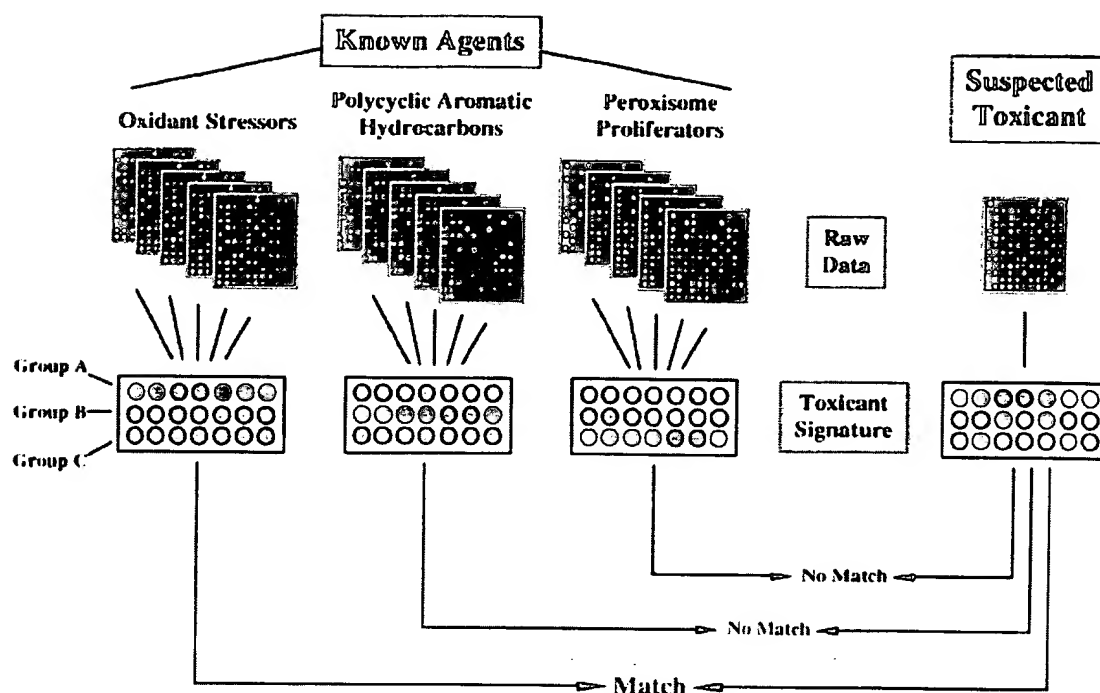


Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce

consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures

are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing ToxChip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

#### Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown



Tabl 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

\*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive *in vivo* test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia*, and *Arabidopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

#### Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.



### Allele s, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

### FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under

which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

### ACKNOWLEDGMENTS

The authors would like to thank Drs. Robert Maronpot, George Lucier, Scott Masten, Nigel Walker, Raymond Tennant, and Ms. Theodora Deverenux for critical review of this manuscript. EFN was supported in part by NIEHS Training Grant #ES07017-24.

### REFERENCES

1. <http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
2. <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 1995;269:496-512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. *Science* 1996;274:546, 563-567.
5. <http://www.perkin-elmer.com/press/prc5448.html>
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* 1992;257:967-971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res* 1996;6:492-503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. *Gene* 1995;156:207-213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484-487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. *Science* 1995;270:467-470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 1996;14:457-460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Biotechnol* 1997;15:1359-1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. *Nat Biotechnol* 1998;16:27-31.
14. <http://www.synteni.com>
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;6:639-645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Biomedical Optics* 1997;2:364-374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998;58:5009-5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA* 1996; 93:10614-10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997;94:13057-13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc Natl Acad Sci USA* 1997;94:2150-2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680-686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 1996;37:29-40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the *E. coli* genome. *Genomics* 1996;37:77-86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *Biotechniques* 1994;17:328-329, 332-336.
25. <http://www.resgen.com/>
26. <http://www.genomesystems.com/>
27. <http://www.clontech.com/>
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA* 1994;91:5022-5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991;251:767-773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA* 1996;93:13555-13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995;19:442-447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675-1680.
34. <http://www.mdyn.com/>
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 1996;33:445-456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. *Science* 1996;274:610-614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. *Nat Genet* 1998;18:155-158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Hum Mutat* 1996;7:244-255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat Genet* 1996;14:441-447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 1996;2:753-759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. *Science* 1998;281:1194-1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. *Environ Health Perspect* 1990;86:313-321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. *Nat Genet* 1998;20:19-23.
45. <http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html>
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 1996;382:722-725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. *J Natl Cancer Inst* 1993;85:1159-1164.
48. <http://www.niehs.nih.gov/envgenom/home.html>



## Expression profiling in toxicology — potentials and limitations

Sandra Steiner \*, N. Leigh Anderson

*Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA*

### Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

**Keywords:** Proteomics; Genomics; Toxicology

### 1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the

pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

\* Corresponding author. Tel.: +1-301-4245989; fax: +1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

## 2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

## 3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

## 4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

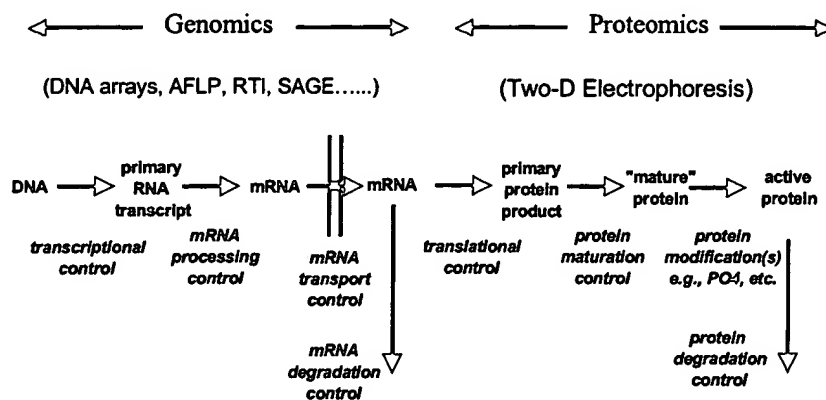


Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.



cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

## 6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

## 7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

## References

- Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* 19, 1998–2003.
- Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537.
- Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* 12, 907–930.
- Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell. Biochem. Suppl.* 22, 108–116.
- Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* 137, 75–89.
- Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* 63, 2243–2250.

- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. *Science* 274, 610–614.
- Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. *Electrophoresis* 19, 355–363.
- Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773.
- Mann, M., Hojrup, P., Roepstorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* 22, 338–345.
- Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. *Electrophoresis* 14, 157–161.
- Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 251, 467–470.
- Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645.
- Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Raynackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. *Biochem. Biophys. Res. Commun.* 218, 777–782.
- Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. *Biochem. Pharmacol.* 51, 253–258.
- Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. *Curr. Biol.* 6, 1543–1544.

## Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. **Key words:** DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681–685 (1999). [Online 6 July 1999]  
<http://ehpnet1.niehs.nih.gov/docs/1999/107/p681-685rockett/abstract.html>

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (1). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer

has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

## Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in

a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (2). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

## Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

Address correspondence to J. Rockett, Reproductive Toxicology Division (MD-72), National Health and Environmental Effects Research Laboratory, U.S. EPA, Research Triangle Park, NC 27711 USA. Telephone: (919) 541-2678. Fax: (919) 541-4017. E-mail: rockett.john@epa.gov

The authors thank R. Kavlock for envisioning the application of array technology to toxicology at the U.S. Environmental Protection Agency. We also thank T. Wall and B. Deitz for administrative assistance.

This document has been reviewed in accordance with EPA policy and approved for publication. Mention of companies, trade names, or products does not signify endorsement of such by the EPA.

Received 23 March 1999; accepted 22 April 1999.



coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic Microsystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrays, as described by Brown (3). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of  $> 2,500$  spots/cm<sup>2</sup> may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental

arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors affecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

### Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., <sup>32</sup>P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosphorimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (4). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA<sup>+</sup> RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After

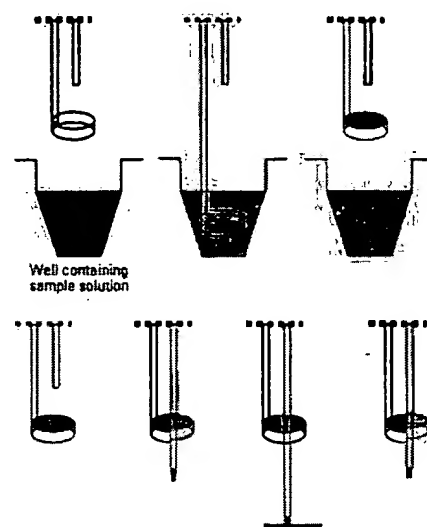


Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (14), with permission from Genetic Microsystems.

normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5' end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

## Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain  $> 10^8$  molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of  $< 1$  fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

## Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (5). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, C.

Table 1. Advantages and disadvantages of different microarray scanning systems.

Nonconfocal laser scanner			
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples		May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
Resolution typically low			

CCD, charge-coupled device.  
From Kawasaki (13).

*elegans* knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes

transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

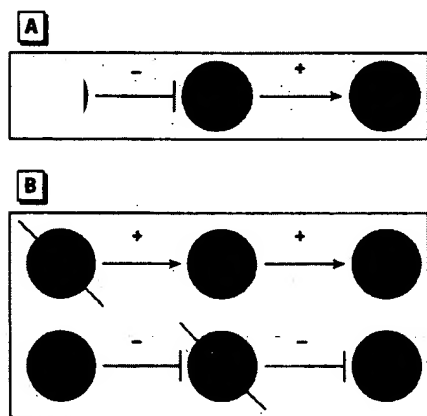
DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

- Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.



**Figure 2.** Potential effects of gene knockout within positively and negatively regulated gene expression networks.  $i_1$  is limiting in wild type for expression of  $i_2$ . (A) A simple, two-component, linear regulatory network operating on gene  $i_2$ , where  $i_1$  is a positive effector of  $i_2$  and  $j_n$  is either a positive or negative effector of  $i_1$ . This network could be deduced by examining the consequence of (B) deleting  $j_n$  on the expression of  $i_1$  and  $i_2$ , where the expression of  $i_2$  would be decreased or increased depending on whether  $j_n$  was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

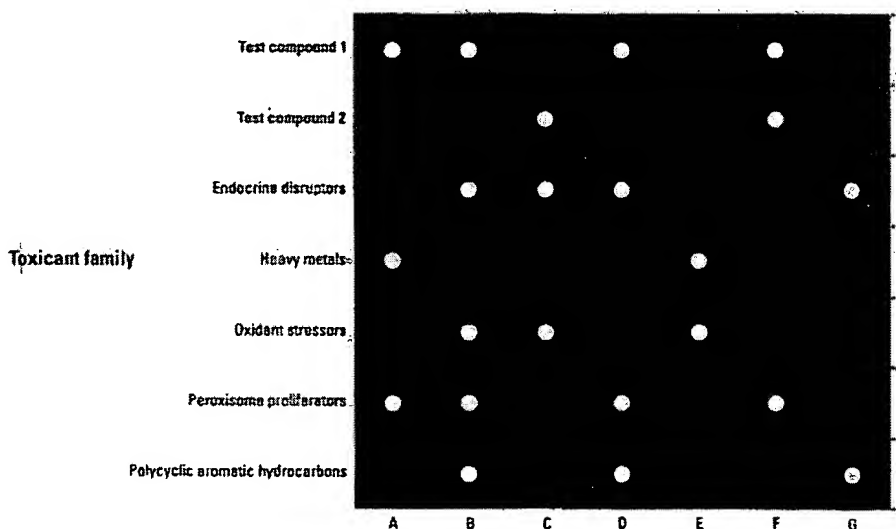
- Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, non-contaminated, sequence-verified, species/cell/tissue/field-specific clones.
- Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
- Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
- Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
- Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
- Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings.

Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.

- Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
- Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
- Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

## EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or



**Figure 3.** Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

## Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene *X* is related to the expression of gene *Y*, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach

is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

## REFERENCES AND NOTES

1. The chipping forecast. *Nat Genet* 21(Suppl 1):3-60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: [www.ncbi.nlm.nih.gov/Schuler/UniGene](http://www.ncbi.nlm.nih.gov/Schuler/UniGene) [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: <http://cmgm.stanford.edu/pbrown> [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* 51:313-324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: [www.mcb.arizona.edu/wardlab/microarray.html](http://www.mcb.arizona.edu/wardlab/microarray.html) [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* 4:1293-1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: <http://cmgm.stanford.edu/pbrown/yeastchip.html> [cited 22 March 1999].
8. Nuwaysir EF, Bitner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. *Mol Carcinog* 24(3):153-159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. *Bioessays* 20:555-561 (1998).
10. Zacharewski TR, Timothy R. Zacharewski. Available: [www.bchmsu.edu/faculty/zachar.htm](http://www.bchmsu.edu/faculty/zachar.htm) [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. *Mol Hum Reprod* 3:473-478 (1997).
12. Stipp D. Gene chip breakthrough. *Fortune*, March 31:56-73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: <http://www.geneticmicro.com/resources/html/coldspring.html> [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

## SPEAKERS

Cindy Afshari  
NIEHS  
Linda Birnbaum  
U.S. EPA  
Ron Butow  
University of Texas  
Southwestern Medical  
Center  
Alex Chenchik  
Clontech Laboratories, Inc.  
David Dix  
U.S. EPA

Abdel Elkahoul  
Research Genetics, Inc.  
Sue Fenton  
U.S. EPA  
Norman Hecht  
University of Pennsylvania  
Pat Hurban  
Paradigm Genetics, Inc.  
Bob Kavlock  
U.S. EPA  
Ernie Kawasaki  
General Scanning, Inc.

Steve Krawetz  
Wayne State University  
Nick Mace  
Genetic Microsystems, Inc.  
Scott Mordecai  
Affymetrix, Inc.  
Kevin Morgan  
Glaxo Wellcome, Inc.  
Elaine Poplin  
Research Genetics, Inc.  
Don Rose  
Cartesian Technologies, Inc.

Jim Samet  
U.S. EPA  
Sam Ward  
University of Arizona  
Jeff Welch  
U.S. EPA  
Reen Wu  
University of California  
at Davis  
Tim Zacharewski  
Michigan State University

**Subject: RE: [Fwd: Toxicology Chip]**

**Date: Mon. 3 Jul 2000 08:09:45 -0400**

**From: "Afshari.Cynthia" <afshari@niehs.nih.gov>**

**To: "Diana Hamlet-Cox" <dianahc@incyte.com>**

You can see the list of clones that we have on our 12K chip at:

<http://manuel.niehs.nih.gov/maps/quest/clonesrch.cfm>

We selected a subset of genes (2000K) that we believed critical to tox response and basic cellular processes and added a set of clones and ESTs to this. We have included a set of control genes (80+) that were selected by the NHGRI because they did not change across a large set of array experiments. However, we have found that some of these genes change significantly after tox treatments and are in the process of looking at the variation of each of these 80+ genes across our experiments.

Our chips are constantly changing and being updated and we hope that our data will lead us to what the toxchip should really be.

I hope this answers your question.

Cindy Afshari

> -----

> From: Diana Hamlet-Cox

> Sent: Monday, June 26, 2000 8:52 PM

> To: afshari@niehs.nih.gov

> Subject: [Fwd: Toxicology Chip]

>

> Dear Dr. Afshari,

>

> Since I have not yet had a response from Bill Grigg, perhaps he was not the right person to contact.

>

> Can you help me in this matter? I don't need to know the sequences,

> necessarily, but I would like very much to know what types of sequences

> are being used, e.g., GPCRs (more specific?), ion channels, etc.

>

> Diana Hamlet-Cox

>

> ----- Original Message -----

> Subject: Toxicology Chip

> Date: Mon. 19 Jun 2000 18:31:48 -0700

> From: Diana Hamlet-Cox <dianahc@incyte.com>

> Organization: Incyte Pharmaceuticals

> To: grigg@niehs.nih.gov

>

> Dear Colleague:

>

> I am doing literature research on the use of expressed genes as

> pharmacotoxicology markers, and found the Press Release dated February

> 29, 2000 regarding the work of the NIEHS in this area. I would like to

> know if there is a resource I can access (or you could provide?) that

> would give me a list of the 12,000 genes that are on your Human ToxChip

> Microarray. In particular, I am interested in the criteria used to

> select sequences for the ToxChip, including any control sequences

> included in the microarray.

>

> Thank you for your assistance in this request.

>

> Diana Hamlet-Cox, Ph.D.

> Incyte Genomics, Inc.

>

> --

>

> =====

## Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER<sup>\*†‡</sup>, CYRUS CHOTHIA<sup>\*</sup>, AND TIM J. P. HUBBARD<sup>§</sup>

<sup>\*</sup>MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and <sup>§</sup>Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

**ABSTRACT** Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA  $k_{\text{tup}} = 1$ , and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests have evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

**Previous Assessments of Sequence Comparison.** Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith-Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed ( $k_{\text{tup}} = 2$ ) or greater effectiveness ( $k_{\text{tup}} = 1$ ). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

Abbreviation: EPO, errors per query.

<sup>†</sup>Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

<sup>‡</sup>To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6\$2.00/0  
PNAS is available online at <http://www.pnas.org>.



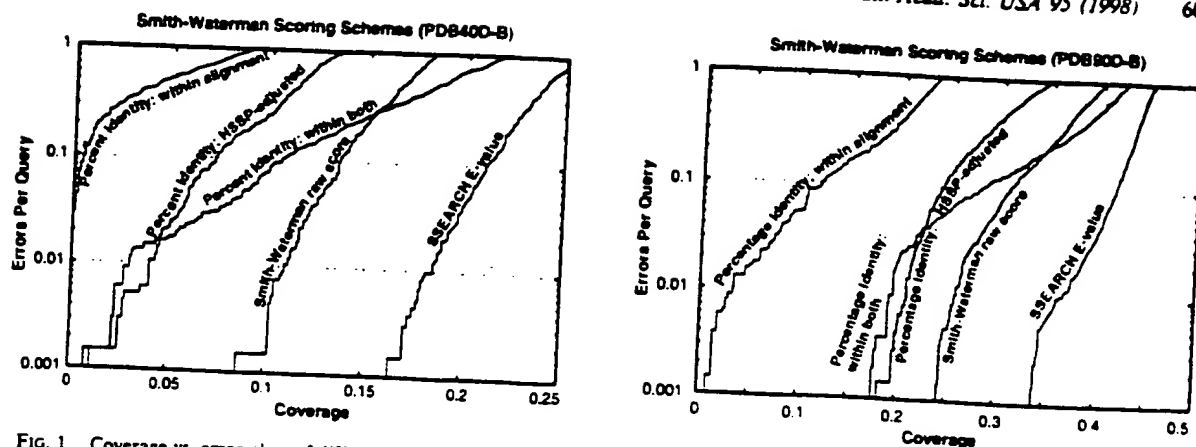


FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith-Waterman. (A) Analysis of PDB400-B database. (B) Analysis of PDB900-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB400-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB400-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues  $l$  is length for  $10 < l < 80$ ;  $H > 100$  for  $l < 10$ ;  $H = 24.7$  for  $l > 80$ . The percentage identity HSP-adjusted score is the percent identity within the alignment minus  $H$ . Smith-Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how

protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Receiver Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

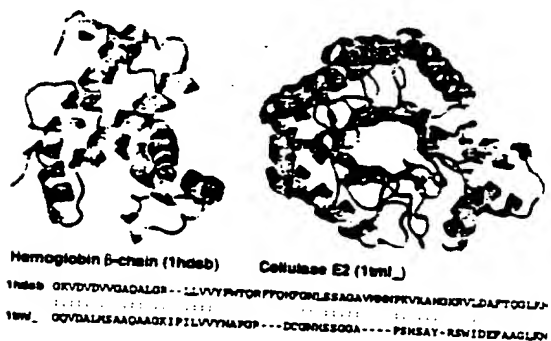


FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin  $\beta$ -chain (PDB code 1hds chain b, ref. 38, Left) and cellulase E2 (PDB code 1tml, ref. 39, Right) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

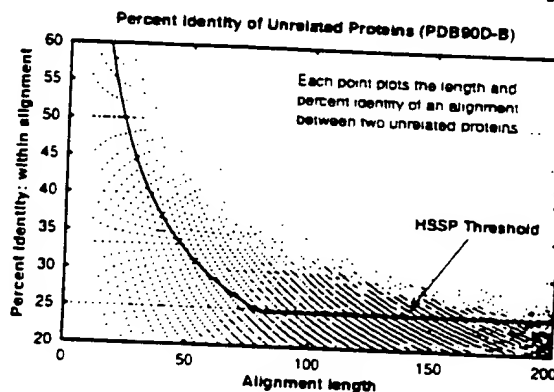


FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB900-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSP threshold (though it is intended to be applied with a different matrix and parameters).

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPO for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPO.

**Overall Detection of Homologs and Comparison of Algorithms.** The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPO. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA ktup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

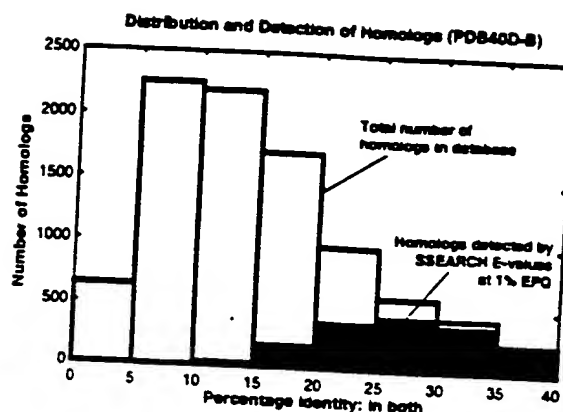


FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPO. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

## CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPO Cutoff	Coverage at 1% EPO
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP = 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

\*Times are from large database searches with genome proteins.



# Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

**A**mong the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should pro-

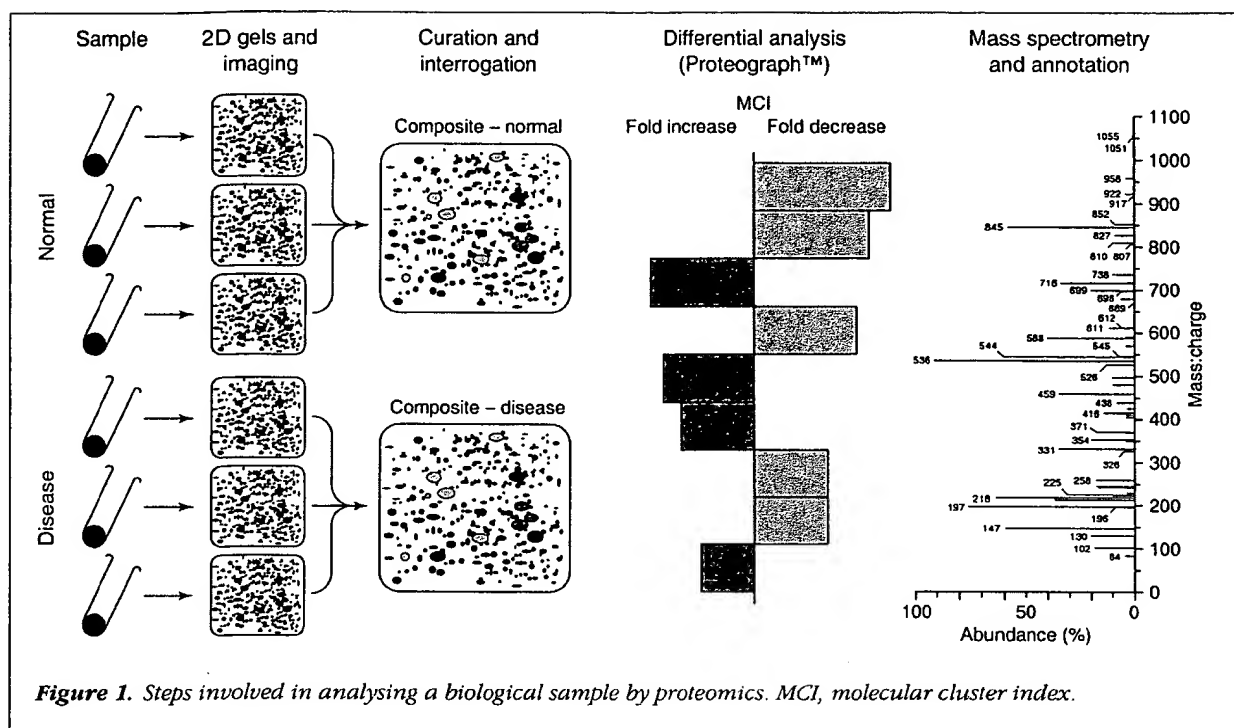
duce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic<sup>1,2</sup> and microarray<sup>3,4</sup> technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

---

**Martin J. Page\*, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. \*tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk



analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

#### Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed<sup>5-7</sup>. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins<sup>8</sup>, which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the

process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS-PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

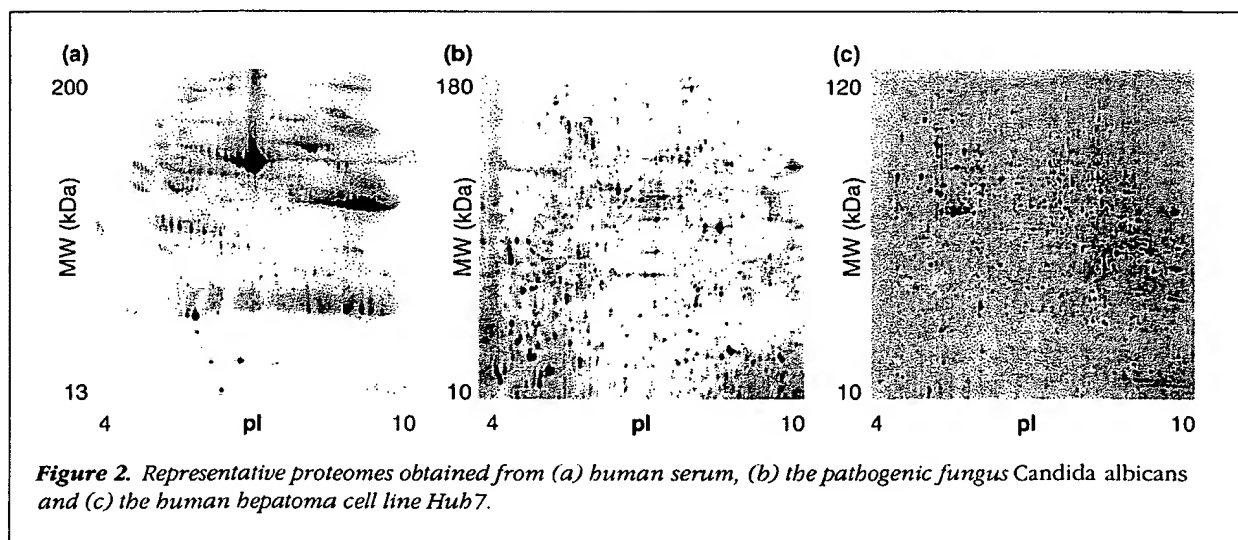
#### Use of proteomics to identify disease specific proteins

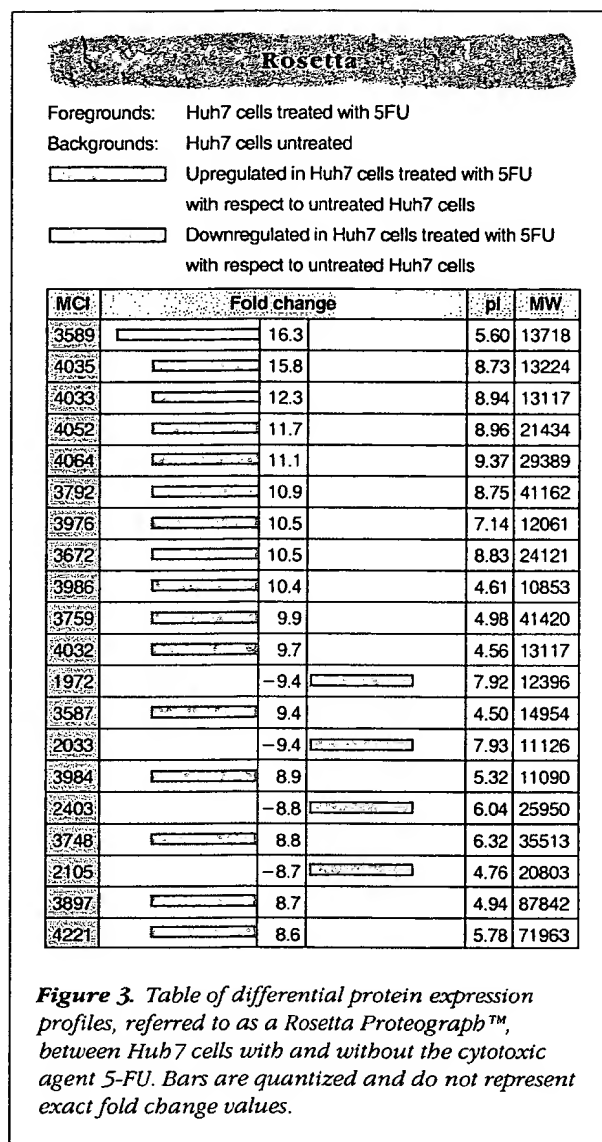
In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and im-

mediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical





cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput

mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

#### Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry<sup>9</sup>. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas<sup>10</sup>, human breast proteins from normal and tumour sources<sup>11–13</sup>, lung tumours<sup>14</sup>, colon tumours<sup>15</sup> and bladder tumours<sup>16</sup>. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified<sup>17,18</sup>.

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

### **Proteomics for target validation and signal transduction studies**

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences<sup>20</sup>.

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules<sup>21–23</sup>. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

#### *Immunoprecipitation studies*

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics<sup>24,25</sup>. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affin-

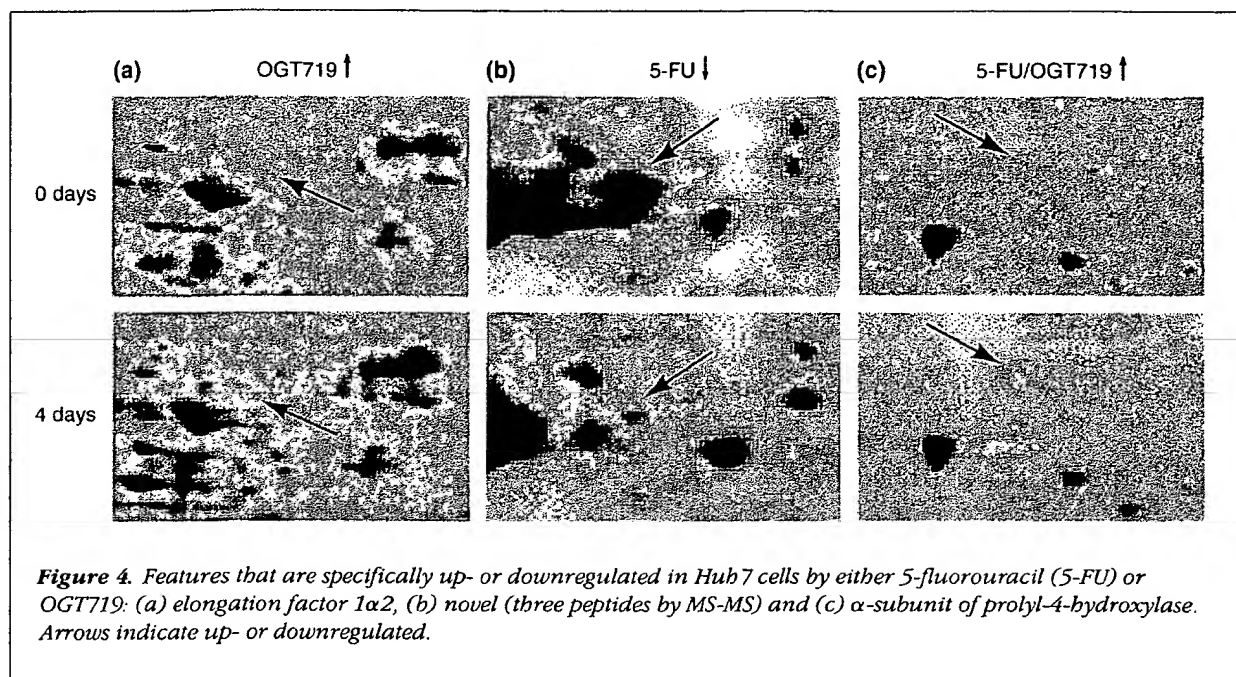
ity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies<sup>26–28</sup>. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

### **Proteomics and drug mode-of-action studies**

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable



of targeting, and being retained in, cells bearing the asialoglycoprotein receptor (ASGP-r), including hepatocytes<sup>29</sup>, hepatoma Huh7 cells<sup>30</sup> and some colorectal tumour cells<sup>31</sup>. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC<sub>50</sub> doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins.

Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein<sup>32</sup>, can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

### *Clear potential*

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

### **Use of proteomics in formal drug toxicology studies**

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from *in vitro* studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems *in vivo*. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an *in vivo* setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members<sup>33,34</sup>, encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabo-

lism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

### *Unique P450 profiles*

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up<sup>35–37</sup>. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.



### Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

### Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

### REFERENCES

- 1 Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
- 2 Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
- 3 Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
- 4 Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
- 5 Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
- 6 James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
- 7 Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
- 8 Parekh, R.B. and Rohlf, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
- 9 Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
- 10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
- 11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
- 12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
- 13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
- 14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
- 15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
- 16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
- 17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
- 18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
- 19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
- 20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
- 21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
- 22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
- 23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
- 24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
- 25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
- 26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
- 27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
- 28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
- 29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
- 30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
- 31 Mu, J.-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
- 32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
- 33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
- 34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
- 35 Vermees, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
- 36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
- 37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210



# Molecular characterization and expression analysis of leucine-rich $\alpha$ 2-glycoprotein, a novel marker of granulocytic differentiation

Lynn C. O'Donnell, Lawrence J. Druhan, and Belinda R. Avalos

*Bone Marrow Transplant Program, The Arthur G. James Cancer Hospital and Richard J. Solove Research Institute, The Ohio State University College of Medicine and Public Health, Columbus*

**Abstract:** Using data obtained from cDNA representational difference analysis to identify genes induced during neutrophilic differentiation of the 32D clone 3G (32Dcl3G) cells, we isolated cDNA clones for murine and human leucine-rich  $\alpha$ 2-glycoprotein (hLRG), a protein with unknown function purified 25 years ago. Expression of LRG during differentiation of 32Dcl3G cells preceded the expression of lactoferrin and gelatinase but followed myeloperoxidase. LRG transcripts were also detected in human neutrophils and progenitor cells but not in peripheral blood mononuclear cells. Notably, LRG expression was up-regulated during neutrophilic differentiation of human MPD and HL-60 cells but down-regulated during monocytic differentiation of HL-60 cells. The hLRG gene was localized to chromosome 19p13.3, a region to which the genes for several neutrophil granule enzymes also map. The putative promoter region of LRG was found to contain consensus-binding sites for PU.1, C/EBP, STAT, and MZF1. These results suggest that LRG is a novel marker for early neutrophilic granulocyte differentiation. *J. Leukoc. Biol.* 72: 478–485; 2002.

**Key Words:** RDA · 32Dcl3G · LRG · myelopoiesis

## INTRODUCTION

Leucine-rich  $\alpha$ 2-glycoprotein (LRG) was first identified as a trace protein in human serum in 1977 [1]. The primary sequence of LRG was determined using Edman degradation, which revealed a marked periodicity in the leucine residues in this protein [2]. At least eight repeating 24 amino acid segments with a notable consensus sequence were identified in LRG. This 24 amino acid consensus sequence, termed the leucine-rich repeat (LRR), has since been identified in a large family of proteins [3, 4]. Although the functions of many of the members of the LRR-containing superfamily are known, the function of LRG has not been elucidated.

Considerable efforts have focused on the mechanisms by which granulocyte-colony stimulating factor (G-CSF) induces granulocytic differentiation [5]. The murine interleukin (mIL)-3-dependent 32D clone 3G (32Dcl3G) cell line was established

from normal murine diploid bone marrow cells and differentiates into mature neutrophils in response to G-CSF [6]. This cell line has been used as a model system for investigating G-CSF-induced differentiation. To identify genes induced by G-CSF during neutrophilic granulocyte differentiation, we used cDNA representational difference analysis (RDA) to generate a representational cDNA library enriched for neutrophil-specific transcripts from G-CSF-treated 32Dcl3G cells. Further characterization of one of the positive clones obtained using this method revealed its identity as LRG. We report here the identification of mLRG and human LRG (hLRG) cDNA clones and the genomic sequences of the LRG genes, the structures of these genes, and their chromosomal localization. Additionally, we show that expression of LRG is induced early during granulocytic differentiation and persists through the neutrophil stage. Expression of LRG in myeloid cells appears to be specific for the neutrophilic granulocyte lineage, as no expression of LRG could be detected in primary human monocytes or cell lines induced to differentiate along the monocyte pathway. G-CSF and chemical inducers of neutrophilic differentiation up-regulated the expression of LRG, indicating that this effect is not unique to the G-CSF pathway. These results suggest that LRG is a novel marker for neutrophilic granulocyte differentiation.

## MATERIALS AND METHODS

Purified recombinant human (rh)G-CSF was generously provided by Amgen Inc. (Thousand Oaks, CA). G-CSF responsive murine 32Dcl3G cells were kindly provided by Dr. Giovanni Rovera (The Wistar Institute, Philadelphia, PA); the human promyelocytic leukemia HL-60 cell line, by Dr. Jas Lang (The Ohio State University, Columbus); the G-CSF responsive human MPD cell line derived from a patient with a myeloproliferative disorder, by Drs. Michael Baumann and Cassandra Paul (Wright State University, Dayton, OH); and WEHI-3B cells, by Dr. Harvey Lodish (Massachusetts Institute of Technology, Cambridge).

### Cells

All cell lines were maintained in RPMI-1640 medium (Gibco-BRL, Grand Island, NY), supplemented with 10% v/v fetal bovine serum, 2 mM glutamine,

---

Correspondence: Dr. Belinda R. Avalos, The Ohio State University, Bone Marrow Transplant Program, A437A Starling-Loving Hall, 320 West Tenth Avenue, Columbus, OH 43210. E-mail: avalos-1@medctr.osu.edu

Received December 20, 2001; revised March 21, 2002; accepted May 11, 2002.

and antibiotics. IL-3-dependent 32Dcl3G cells also required the presence of conditioned medium from WEHI 3B cells (10% v/v) as a source of IL-3 [7]. The 32Dcl3G cells were washed twice with phosphate-buffered saline to remove IL-3 before transfer to medium containing rhG-CSF (10 ng/mL) to induce neutrophilic granulocyte differentiation. Granulocytic differentiation of HL-60 cells was initiated by the addition of dimethyl sulfoxide (DMSO; 1.25% v/v) to the culture medium [8]. For monocytic differentiation of HL-60 cells, phorbol 12-myristate 13-acetate (PMA; 5 nM) was added to the culture medium [9]. For the maintenance of MPD cells, sodium pyruvate was added to the medium to a final concentration of 1 mM. Neutrophilic granulocyte differentiation of MPD cells was initiated by the addition of rhG-CSF (2.85 ng/mL) [10]. Differentiation of all cell lines was confirmed by morphology using Wright-Giemsa staining and reverse transcriptase-polymerase chain reaction (RT-PCR; see below) for expression of the neutrophil markers myeloperoxidase (MPO), lactoferrin (LF), and gelatinase (GEL) or by nitroblue tetrazolium (NBT) assays [11].

Peripheral blood (PB) and bone marrow (BM) were obtained from hematologically normal donors following informed consent and were collected in EDTA-containing tubes. Mononuclear cell fractions were isolated using density gradient centrifugation with Histopaque 1.077 (Sigma Chemical Co., St. Louis, MO). Neutrophils were isolated from the resulting cell pellet using dextran sulfate (3% w/v) sedimentation and hypotonic lysis of red blood cells. Wright-Giemsa staining was used to confirm >97% purity of the neutrophils, which were used for subsequent analysis.

## cDNA RDA

To identify genes induced during G-CSF-stimulated differentiation, RNA was extracted from 32Dcl3G cells grown in the presence of IL-3 or in the presence of rhG-CSF. Trizol™ reagent (Gibco-BRL) was used to extract total RNA from four different populations of 32Dcl3G cells: those grown in the presence of IL-3 for 5 days (designated I5) and 7 days (I7) and those grown in the presence of rhG-CSF for 5 days (G5) and 7 days (G7). Poly-A RNA was isolated from each total RNA preparation using the PolyATtract system (Promega, Madison, WI). The resultant mRNA preparations were treated with DNase I, phenol/chloroform-extracted, ethanol-precipitated, and resuspended in water. For the preparation of cDNA, 5 µg from the 5 day and 7 day mRNA preparations for each cytokine (I5+I7 and G5+G7) were pooled and diluted to 0.4 µg/µL. The pooled samples were then used to produce double-stranded cDNA using the cDNA Synthesis System kit (Gibco-BRL) according to the manufacturer's instructions.

In preparation for cDNA RDA, the double-stranded cDNAs were precipitated twice at -20°C with 1/5 vol 10 M ammonium acetate and an equal volume of isopropanol and were resuspended in TE. Using the technique described by Hubank and Schatz [12], we used 2 µg each pooled cDNA in the preparation of separate driver and tester representational cDNA populations. The driver cDNA was from IL-3-treated cells (I5+I7), and the tester cDNA was from G-CSF-treated cells (G5+G7). RDA was then performed using driver cDNA hybridized in excess against tester cDNA under conditions where subsequent PCR reactions subtracted sequences common to both populations and amplified sequences unique to the tester.

After three rounds of hybridization and subtraction, the heterogeneous final difference product (DP3) was digested with *DpnII*, cloned into the *Bam*HI site of the pBluescript KS+ phagemid (Stratagene, San Diego, CA), and transformed into DH5α *Escherichia coli*. Each clone contained a *DpnII* cDNA fragment from a population enriched for G-CSF-induced transcripts. cDNA inserts from individual clones were then PCR-amplified and screened for differential expression by dot blot analysis using gel-purified driver DP3 and tester DP3 as probes. This initial screen confirmed that 76 of 141 RDA clones contained cDNA inserts enriched in the tester DP3. For clones that were found to be more abundant in the tester DP3 by dot blot, differential expression was subsequently confirmed by Northern blot analysis of the original total RNA samples. The differentially expressed RDA clones were then sequenced and compared with the public nucleotide and protein databases using BLAST [13].

## Identification and characterization of cDNAs and genomic sequences corresponding to differentially expressed RDA clones

BLAST queries of the mouse and human expressed sequence tag (EST) databases were used to identify putative full-length cDNA clones of differen-

tially expressed RDA clones. Additional cDNA clones were identified in UniGene clusters that contained the EST clones of interest. Candidate EST clones were purchased from Genome Systems Inc. (Incyte Genomics, Inc., Palo Alto, CA) or Research Genetics Inc. (Invitrogen, Carlsbad, CA) and were analyzed by restriction analysis and end-sequencing. Both strands of the clones most likely to contain the full-length cDNA of interest were sequenced. The cDNA sequences were scanned for open reading frames (ORF) using MacVector (Eastman Kodak, Rochester, NY). The putative ORFs were analyzed for the presence of signal peptides using the algorithms SignalP V2.0 and TargetP V1.0 [14–16].

Genomic sequences corresponding to the identified full-length cDNA clones were identified by BLAST queries of GenBank, the mouse genome database, or the Celera Inc. (Rockville, MD) human genome database [17–19]. Chromosomal assignments were based on the published physical location of the identified genomic clone or by analysis of the genomic sequence with electronic PCR [20]. The intron/exo structures of the identified genomic sequences were determined by alignment of the full-length cDNA to the genomic sequence using BLAST2 [21]. Further analysis of the genomic sequences using McPromoter V3b and the Berkeley Drosophila Genome Project (BDGP) Neural Network Promoter Prediction algorithms (University of California, Berkeley) was done to identify putative transcription start sites [22, 23]. Putative promoter regions were scanned against the TRANSFAC database using MatInspector Professional (Genomatix, München, Germany) to identify putative binding sites for transcription factors [24, 25]. Genomic sequences were compared within the 1000-bp region upstream of each of the identified ORFs by Pustell DNA dot matrix analysis using a window-sized setting of 20, minimum homology score of 65%, and hash value of 4 [26].

## Northern blot hybridization

Total RNAs (10 µg each) were irreversibly denatured with glyoxal/DMSO and were then size-fractionated on a 1% w/v agarose gel [27]. RNA was transferred to charged nylon membranes and hybridized with murine- or human-specific probes. A 537-bp murine-specific LRG probe was prepared by liberation of the insert from the G11 RDA clone with *Eco*RI and *Nco*I. A 395-bp, human-specific LRG probe was separated from IMAGE clone 81861 that had been digested with *Bst*XI and *Eco*RI. This probe is common to both classes of hLRG transcripts. To prepare a class II-specific hLRG probe, the 3' end of IMAGE clone 85213 was PCR-amplified using a primer specific for the 3' end of the class II LRG transcript (5'-GCTTCCTAGAACACACCGATG-3') and a primer specific for the LacZ portion of the vector (5'-CCCAGTCACGACGTTGTAAACG-3'). The amplified product was digested with *Xho*I to remove the vector sequence, yielding a 295-bp fragment containing sequence specific for the class II hLRG. All probes were gel-purified prior to labeling using the QIAquick gel extraction kit (Qiagen, Chatsworth, CA). Probes were labeled with <sup>32</sup>P using the High Prime kit (Roche/BMB, Nutley, NJ) and were purified using size-exclusion spin columns. Hybridization and washing were performed with ExpressHyb (Clontech, Palo Alto, CA) hybridization solution according to the manufacturer's instructions. Blots were exposed to X-ray film or a phosphor-imaging screen (Molecular Dynamics, Sunnyvale, CA). Northern blot quantitation was performed using the Imagequant software (Molecular Dynamics). The mouse multiple tissue Northern blot used was purchased from Clontech.

## RT-PCR

Reverse transcription was carried out on 2 µg total RNA from each sample using Superscript™ RT (Gibco-BRL) and random hexamer primers according to the manufacturer's instructions. PCR with gene-specific primers was performed using *Taq* DNA polymerase (Gibco-BRL) at a final MgCl<sub>2</sub> concentration of 2 mM. The PCR products were size-fractionated on agarose/TAE gels (1% w/v) and visualized with ethidium bromide staining. The thermocycling program for all reactions included an initial incubation at 94°C for 5 min, followed by 30 cycles of denaturation (94°C, 30 s), annealing (see temperatures below, 30 s), and extension (72°C, 30 s), with a final incubation at 72°C for 7 min. PCR with the mMPO-specific primers mPofor (5'-AACCAGCTGGGGCTGCTGGCTGCAATACAG-3') and mMPrev (5'-AACTCCAGGTTCITCAGCACCGTCCG-3') was performed at an annealing temperature of 62°C, which resulted in an 806-bp product. PCR with the murine LF-specific primers mLFfor (5'-GCCAGTCACAGGAGAAGTTTGG-3') and mLFrev (5'-GCCATTGCTTTTGAGGATTTC-3') was performed at an annealing temper-

ature of 54°C, resulting in a 452-bp product. PCR with the murine GEL-specific primers mGEL<sub>for</sub> (5'-ACAACTGAACCACAGCCGACAG-3') and mGEL<sub>rev</sub> (5'-TCATTTTGGAAACTCACAGCC-3') was performed with an annealing temperature of 54°C, producing a 743-bp product.

## RESULTS

### Isolation of differentially expressed LRG clones

We initially examined the time course for G-CSF-induced morphologic changes in 32Dcl3G cells to optimize our chances of isolating genes that are induced early during neutrophilic granulocyte differentiation. G-CSF treatment of 32Dcl3G cells resulted in a decrease in the blast cell population from 90% at day 0 to 10% at day 14 (Fig. 1). The presence of mature neutrophils could be detected by day 5, at a time when more than 50% of the cells had differentiated into myelocytes and metamyelocytes. With continued growth in G-CSF, the numbers of myelocytes and metamyelocytes peaked by day 7, as neutrophil numbers continued to increase. Based on these initial experiments, RNA was extracted from 32Dcl3G cells grown in the presence of G-CSF for 5 days and 7 days. The RDA analysis compared the pooled RNA samples from the G-CSF-treated cells against pooled RNA samples from cells grown in the absence of G-CSF (but in the presence of IL-3) for an equal number of days. This procedure yielded multiple clones, which were confirmed to be differentially expressed in G-CSF-treated cells by Northern blot analysis of the pooled RNA samples (data not shown). BLAST queries of the deduced amino acid sequences for two of the clones isolated (designated G8 and G11), indicated that both clones had significant homology to different portions of the published sequence for the hLRG protein (Fig. 2). The 121-bp G8 clone showed signifi-

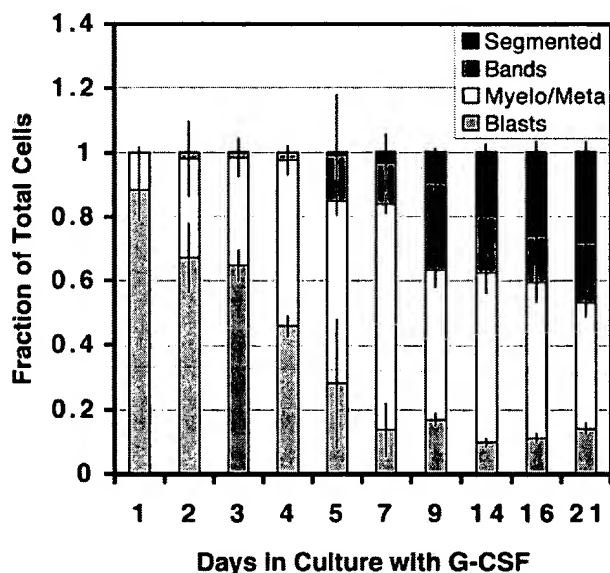


Fig. 1. Time course of morphologic changes associated with G-CSF-induced neutrophilic differentiation of 32Dcl3G cells. Cells were grown in the presence of G-CSF (10 ng/mL) for the indicated times, cytopun, then stained with Wright-Giemsa solution, and examined by light microscopy. At each time point, a differential cell count on 100 cells was performed. Data are the averages from three independent experiments.

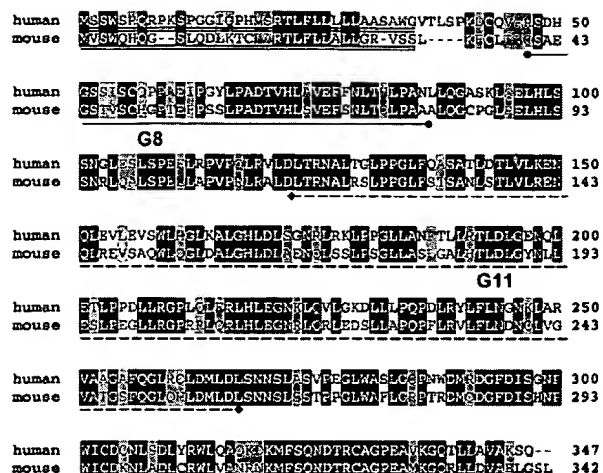


Fig. 2. Alignment of deduced amino acid sequences for hLRG and mLRG, indicating significant homology. The deduced amino acid sequences from the mLRG and hLRG cDNAs are shown and their alignment, as determined by ClustalW (version 1.81). Black and gray shading indicate identical and homologous amino acids, respectively. Sequences corresponding to the original G8 (●—●) and G11 (◆—◆) clones identified by RDA are indicated. Predicted signal peptides are shown by the double underlines. The predicted mature peptide encoded by the hLRG cDNA is identical to the previously published sequence for hLRG that was purified from serum.

cant homology over a 40 amino acid stretch with the published sequence for the 312 amino acid hLRG protein (F11 to N51). The deduced amino acid sequence of the 436-bp G11 clone was found to be homologous to a more distal stretch of 145 amino acids (D87 to D231) in hLRG. Subsequent sequence analyses indicated that 20% of the differentially expressed clones were LRG. Differentially expressed RDA clones corresponding to MPO, migration-inhibitory factor-related protein 14, and ras-related nuclear protein were also identified.

### Identification, sequencing, and analysis of mLRG and hLRG cDNAs

As RDA does not yield full-length cDNA clones, BLAST queries of the murine and human databases were performed to identify EST clones with homology to the murine RDA clones. Three murine IMAGE clones with homology to the G8 and G11 RDA clones were purchased and characterized. Sequence analysis of the 3' and 5' ends of each of the clones and size determination of the cDNA inserts by restriction analysis suggested that one clone (IMAGE clone ID: 597689) most likely contained a full-length cDNA for mLRG. Both strands of the 1310-bp cDNA insert of this clone were sequenced, and sequence analysis identified a Kozak consensus sequence, a 342 amino acid ORF, a 276-bp 3' untranslated region (UTR), and a poly-A tail (GenBank accession: AF403429). Except for 3 bp upstream from the putative translation start site, no 5' UTR was present. The deduced amino acid sequence of the ORF shares 66% identity and 76% homology with the hLRG protein sequence (Fig. 2). SignalP V2.0 predicts a signal peptide at amino acids 1–32 (signal peptide probability=0.967, signal anchor probability=0.000, and a maximum cleavage site probability at residue 33 of 0.69). Analysis of the sequence by

TargetP V1.0 suggests that mLRG is targeted to the secretory pathway.

BLAST searches of the human EST database, along with analysis of the UniGene cluster (Hs10844), identified several EST clones with homology to the LRG protein sequence. Alignment of these ESTs indicated the existence of two possible classes of LRG transcripts (designated classes I and II) differing only in the size of the 3' UTR. IMAGE clones with the largest cDNA inserts from each class were sequenced. The class I clone (IMAGE clone ID: 2426875), which was 1261-bp in length, contained a very small 5' UTR, an ORF that encoded a putative signal peptide, followed by the entire amino acid sequence of hLRG (as determined by Takahashi et al. [2]), and a 191-bp 3' UTR that contained an Alu repetitive element extending from bp 1133 to 1192.

The class II clone (IMAGE clone ID: 2403704) was identical to the class I clone except for a much longer 726-bp 3' UTR (GenBank accession: AF403428). Analysis of the deduced amino acid sequence of the hLRG coding region predicted a signal peptide at amino acids 1–35 (signal peptide probability=0.992, signal anchor probability=0.008, and maximum cleavage site probability at residue 36 of 0.597). Similar to mLRG, TargetP V1.0 suggests that hLRG is a secreted protein.

On the basis of these results, we have identified one mLRG and two hLRG cDNA clones. The mLRG cDNA is 1310-bp in length, and the two hLRG cDNAs are 1261-bp and 1801-bp in length.

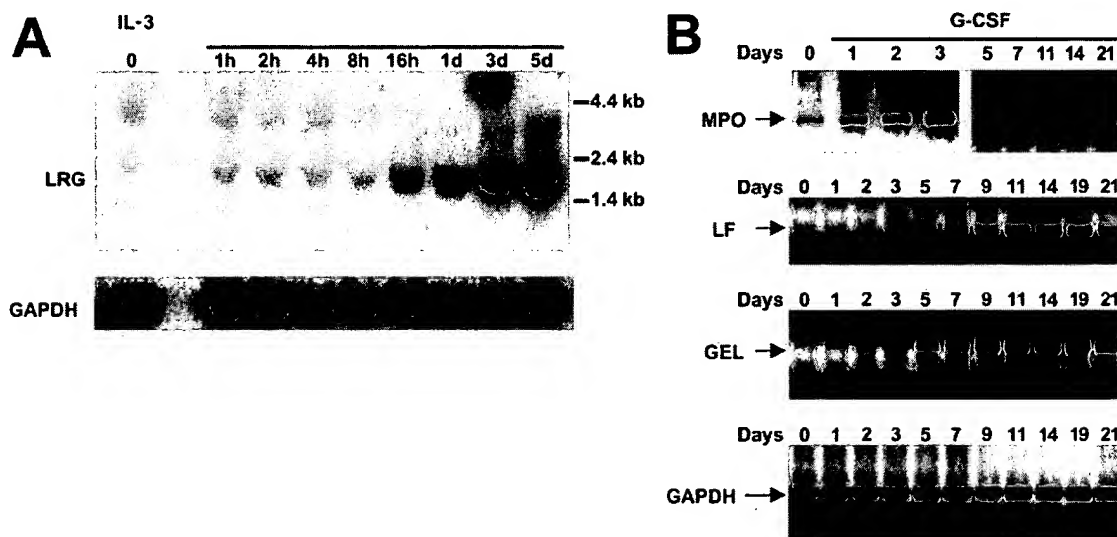
### Analysis of LRG expression

Induction of expression of LRG during G-CSF-induced granulocytic differentiation of 32Dcl3G cells was analyzed by Northern blotting (Fig. 3A). An approximate 1.6-kb transcript could be detected as early as 16 h after the addition of G-CSF

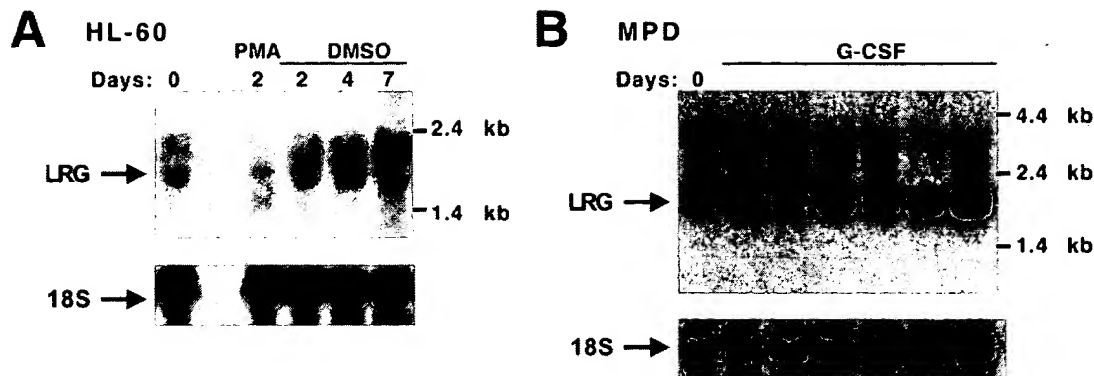
to the culture medium. Expression of this transcript steadily increased thereafter, with an 80-fold increase at 5 days.

We next investigated the time course for LRG expression relative to other genes known to be induced and/or up-regulated during neutrophilic granulocyte differentiation (Fig. 3B). MPO was constitutively expressed in 32Dcl3G cells, whereas no transcripts for LF or GEL could be detected in untreated 32Dcl3G cells. G-CSF treatment of 32Dcl3G cells resulted in up-regulation of MPO expression and also in the induction of expression of LF and GEL at 3 days and 5 days, respectively. On the basis of these findings, we conclude that expression of LRG follows the expression of MPO in 32Dcl3G cells but precedes the expression of LF and GEL. These results suggest that the expression of LRG is a relatively early event in neutrophilic granulocyte differentiation.

We also examined the expression of LRG in two human myeloid cell lines that differentiate along the neutrophilic granulocyte pathway. Treatment of the human promyelocytic leukemia cell line HL-60 with DMSO, which induces granulocytic differentiation [9], resulted in a twofold increase in the level of the 1.8-kb LRG transcript by day 7 (Fig. 4A). Neutrophilic granulocyte differentiation of the HL-60 cells was confirmed morphologically by Wright-Giemsa staining and the NBT assay. In contrast, treatment with PMA, which induces monocytic differentiation of HL-60 cells [28], resulted in a decrease in LRG expression. The low level of basal expression of LRG detected in untreated HL-60 cells may reflect a more mature phenotype of HL-60 cells, as compared with 32Dcl3G cells. The observation that DMSO-induced granulocytic differentiation of HL-60 cells is associated with up-regulation of LRG expression demonstrates that this effect is not unique to G-CSF.



**Fig. 3.** LRG expression is induced during neutrophilic differentiation of 32Dcl3G cells and precedes expression of the secondary and tertiary granule enzymes LF and GEL but follows MPO. 32Dcl3G cells were grown in the presence of G-CSF (10 ng/ml) for the indicated times and RNA-extracted. (A) Northern blot analysis using a probe specific for mLRG. For each time point, 10  $\mu$ g total RNA was loaded. RNA molecular weight markers are shown to the right. In the lower panel, the blot in the upper panel was stripped and reprobed with a glyceraldehyde 3-phosphate dehydrogenase (GAPDH)-specific probe to control for RNA loading. (B) RT-PCR analysis of the expression of MPO, LF, GEL, and GAPDH. The PCR products were size-fractionated on agarose/TAE gels (1% w/v) and visualized with ethidium bromide staining.



**Fig. 4.** LRG expression is up-regulated during neutrophilic differentiation of HL-60 and MPD cells but down-regulated during monocytic differentiation. (A) HL-60 cells treated with PMA (5 nM) or DMSO (1.25% v/v) to induce monocytic or neutrophilic differentiation, respectively. (B) MPD cells treated with G-CSF (2.8 ng/ml) to induce neutrophilic differentiation. At each time point, RNA was extracted for analysis by Northern blotting using a probe for hLRG. For each time point, 10  $\mu$ g total RNA was loaded. Following hybridization with the LRG probe, each blot was stripped and reprobbed with an 18S rRNA-specific probe (lower panels). Molecular weight markers are shown to the right. Treatment of HL-60 cells was carried out for 2 days, at which time the cells became adherent consistent with a monocyte/macrophage phenotype.

We also observed up-regulation of LRG expression during neutrophilic granulocyte differentiation in the human MPD cell line, which differentiates in response to G-CSF, and expression of proteins belongs to all three classes of neutrophilic granules [10]. Similar to HL-60 cells, LRG expression was also detected in untreated MPD cells. G-CSF treatment of MPD cells resulted in an increase in expression of the 1.8-kb LRG transcript, which could be detected as early as 24 h after the addition of rhG-CSF to the culture medium (Fig. 4B). LRG expression in MPD cells increased steadily thereafter, with an approximate threefold increase by day 14. In HL-60 and MPD cells, expression of only the class II LRG transcript was detected.

We next investigated the expression of LRG in primary cells from healthy volunteer donors. LRG transcripts were detected in the neutrophil fraction and in the progenitor-rich mononuclear cell fraction of BM (Fig. 5). In contrast, LRG expression could only be detected in the neutrophil fraction from PB and not in the mononuclear cell fraction. Similar to the results obtained with cell lines, only class II transcripts for hLRG could be detected in BM and PB.

We also examined the tissue distribution for expression of LRG. Northern blot analysis of mRNA samples extracted from adult mouse tissues demonstrated a high level of LRG expression in liver with a much lower level in heart and minimally detectable expression in spleen and lung. No LRG transcripts

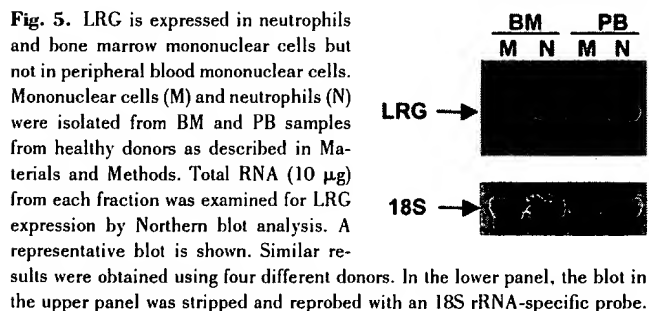
were detected in brain, skeletal muscle, kidney, or testis (data not shown).

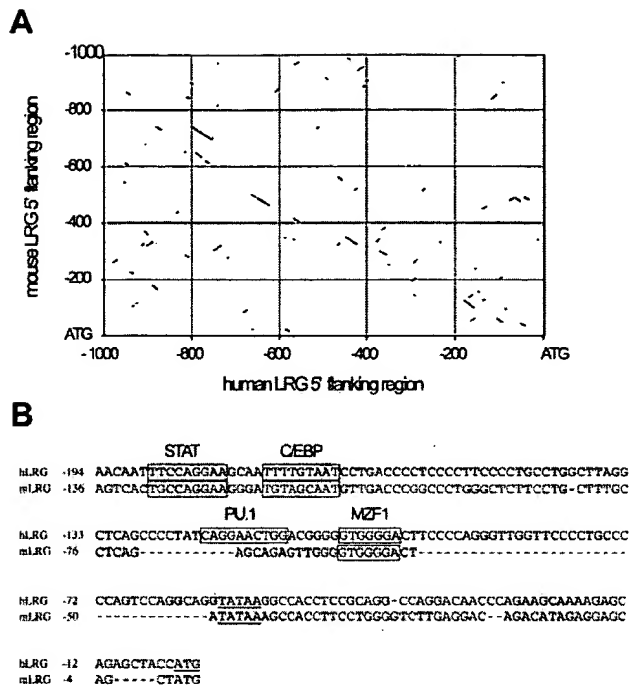
### Chromosomal localization and genomic structure of mLRG and hLRG

A BLAST query of the public Celera Inc. human genome sequence database using the sequence of the human class I clone indicated a 100% match to segment GA xx2HTBKPS83T:1000001:1500000, which localizes to chromosome 19. Electronic PCR of the 6000-bp surrounding the hLRG genomic locus identified two sequence tag sites, stSG44969 and sts-T71373, in the 3' UTR of the hLRG locus. These sequence tag sites belong to the UniGene cluster Hs10844 and have been mapped to 19p13.3 by radiation hybrid mapping [29]. Using BLAST2, alignment of the cDNA sequences for hLRG with the genomic sequence for hLRG indicated that the hLRG gene is composed of two exons. Exon 1 is 56-bp in length and contains 25-bp of the 5' UTR and the first 32-bp of the 1044-bp ORF. Exon 2 is 1737-bp in length, and contains the remainder of the ORF, including the stop codon and the entire 722-bp 3' UTR. The intervening intron is 1030-bp in length and contains consensus splice donor and acceptor sites. Analysis of the genomic sequence upstream from the hLRG coding region indicated a putative transcription start site approximately 30 bp upstream of the ATG. Comparison of the putative promoter regions identified for the hLRG and mLRG genes revealed the presence of clusters of high homology and approximately 60% overall identity (Fig. 6A).

The genomic sequence 500 bp upstream and 100 bp downstream of the putative transcription start site was examined for consensus sequences for DNA binding sites for all transcription factors in the TRANSFAC database [23]. This analysis identified putative binding sites for C/EBP, PU.1, MZF1, and STAT (Fig. 6B).

A BLAST query of the mouse-specific genome database using the mLRG cDNA sequence identified a BAC clone (GenBank accession: AC026385), which mapped to mouse chromosome 11 and contained a portion of the mLRG genomic





**Fig. 6.** Comparison of putative LRG promoter regions of the hLRG and mLRG genes. (A) Pustell DNA dot matrix comparison of the 5' flanking regions. (B) Alignment of human and murine sequences with the putative TATA boxes and the ATG start sites underlined. Boxes denote sequences corresponding to putative binding sites for transcription factors.

sequence, including the putative promoter region, the entire 5' UTR, and a portion of the coding sequence. Alignment of the mLRG cDNA to this BAC indicated the presence of at least two exons in the mLRG genomic locus. Similar to the hLRG gene, exon 1 of the mLRG gene is small at 44 bp in length and contains the first 26 bp of the 1029-bp ORF preceded by a 5' UTR of approximately 18 bp. The intron is 916 bp in length and contains consensus splice-donor and splice-acceptor sites. The sequence of the second exon is incomplete, as the BAC sequence terminates before the end of this exon. Genomic sequence data containing the remainder of the mLRG gene are not currently available. However, the available structure for the mLRG gene is very similar to that of the hLRG gene. Analysis of the genomic sequence upstream of the identified mLRG coding region predicted a transcription start site approximately 30 bp upstream of the ATG. Sequence analysis of the putative mLRG promoter region indicated the presence of putative binding sites for C/EBP, MZF1, and STAT (Fig. 6B).

## DISCUSSION

The leucine-rich repeat family is a diverse family of proteins, which have been shown to be involved in protein-protein interactions, signal transduction, and cell adhesion and development [3]. The function of the founding member of this family, LRG, has remained unknown. Here, we report expression of LRG in primary human neutrophils and the induction and/or

up-regulation of its expression during neutrophilic granulocyte differentiation of human and murine cell lines.

Using cDNA RDA, we identified LRG as being differentially expressed during G-CSF-induced granulocytic differentiation of murine 32Dcl3G cells. BLAST queries using the RDA clones enabled us to obtain full-length cDNA clones for mLRG and hLRG from the EST clone banks. Two classes of hLRG cDNA were identified, which differed only in the length of their 3' UTRs. Comparison of the sequences of the two human mRNA classes with the LRG genomic sequence indicates that the difference is not a result of alternative splicing but rather because of different polyadenylation sites. The polyadenylation site of the class II clone has a canonical AAUAAA signal sequence, but the shorter class I clone does not. It is possible that the class I clone is not a true mRNA species but rather an artificial species produced by the cloning method. As the sequence surrounding the putative polyadenylation site in the class I clone is rich in adenosine bases, mispriming the oligo-dT primer during cDNA synthesis could have given rise to the class I clones. In any event, the only class of LRG transcript detected in cell lines and in primary human cells from PB and BM was the class II transcript.

Expression of LRG in 32Dcl3G cells could be detected as early as 16 h following G-CSF treatment, nearly 4 days prior to the appearance of cells with the morphological features of neutrophils. The time course for induction of LRG expression in relation to other genes known to be up-regulated during granulopoiesis provided additional evidence that LRG expression is induced at an early time point during neutrophilic granulocyte differentiation. LRG expression was found to occur subsequent to the expression of myeloperoxidase, a primary/azurophilic granule protein, but prior to the expression of lactoferrin and gelatinase, which are contained within the secondary and tertiary granules, respectively. In addition, LRG transcripts were detected in the granulocytic and mononuclear cell fractions of bone marrow and in peripheral blood granulocytes but not monocytes or lymphocytes. Collectively, these data suggest that LRG expression during hematopoiesis is specific for differentiation along the neutrophilic granulocyte lineage and that its expression is induced early during this process.

Of interest is the identification of consensus binding sites for myeloid-specific transcription factors in the putative promoter regions of the hLRG and mLRG genes. The C/EBP family of transcription factors has been shown to be involved in early myeloid differentiation [30, 31]. C/EBP $\alpha$ -null mice exhibit a blockade in granulocyte development at the myeloblast stage, and C/EBP $\epsilon$ -null mice were found to have an increase in the numbers of immature myeloid cells in their bone marrow and produced hyposegmented neutrophils that were functionally deficient [32, 33]. An absolute requirement for PU.1 during granulopoiesis has been demonstrated in vitro and in vivo [34, 35]. Two PU.1-deficient mouse strains have been generated. In one case, no viable PU.1 $^{-/-}$  progeny were produced (embryos died by day 18), whereas in the other case, newborn pups died of severe septicemia after 48 h. Both PU.1-null mouse strains were found to be deficient in multiple hematopoietic lineages, including monocytes, neutrophils, B cells, and T cells [36, 37]. Notably, binding sites for PU.1 and C/EBP $\alpha$  have also been

identified in the promoter region for other myeloid-specific genes, including myeloperoxidase, neutrophil elastase, and the G-CSF receptor [38–40].

Consensus binding sites for MZF-1 and STAT were also identified in the putative promoters of the hLRG and mLRG genes. G-CSF-stimulated granulocyte-colony formation has been reported to be inhibited in vitro by antisense oligonucleotides to MZF-1, and STATs have been shown to mediate signaling via many cytokine receptors, including the G-CSF receptor [41, 42].

Results from the TargetP algorithm suggest that LRG is processed along the secretory pathway, consistent with its initial purification from plasma. The SignalP V2.0 algorithm predicts a signal peptide of 35 residues for the human protein and 32 residues for the mouse protein. Although a signal peptide of 35 (or 32) amino acids is unusually long, the predicted cleavage site would result in a protein that is 100% homologous to the primary sequence previously reported for hLRG. The absence of a signal-anchor region, as indicated by the SignalP V2.0 analysis, further supports the notion that LRG is a secreted protein. It is possible that LRG is secreted directly or that it is stored in intracytoplasmic granules prior to secretion, similar to other secreted neutrophil granule proteins such as myeloperoxidase, lactoferrin, and gelatinase. The lack of any currently available antibodies to LRG has hindered the determination of its cellular localization. Future studies using antibody generated against purified LRG as well as epitope-tagged LRG and commercially available antibodies recognizing the tag, which are currently underway in our laboratory, should help to elucidate the cellular localization of LRG and its physiologic role in neutrophilic granulocytes.

## ACKNOWLEDGMENTS

This work was supported by grants CA75226, CA82859, and CA16058 from the National Cancer Institute. We thank Drs. Jas Lang, Giovanni Rovera, Michael Baumann, Cassandra Paul, and Harvey Lodish for generously providing the cell lines used in our studies.

## REFERENCES

- Haupt, H., Baudner, S. (1977) Isolation and characterization of an unknown, leucine-rich 3.1-S- $\alpha$ 2-glycoprotein from human serum. *Hoppe-Seyler's Z. Physiol. Chem.* 358, 639–646.
- Takahashi, N., Takahashi, Y., Putnam, F. W. (1985) Periodicity of leucine and tandem repetition of a 24-amino acid segment in the primary structure of leucine-rich  $\alpha$ 2-glycoprotein of human serum. *Proc. Natl. Acad. Sci. USA* 82, 1906–1910.
- Buchanan, S. G., Gay, N. J. (1996) Structural and functional diversity in the leucine-rich repeat family of proteins. *Prog. Biophys. Mol. Biol.* 65, 1–44.
- Kobe, B., Deisenhofer, J. (1994) The leucine-rich repeat: a versatile binding motif. *TIBS* 19, 415–421.
- Avalos, B. R. (1996) Molecular analysis of the granulocyte colony-stimulating factor receptor. *Blood* 88, 761–777.
- Valtieri, M., Tweardy, D. J., Caracciolo, D., Johnson, K., Mavilio, F., Altmann, S., Santoli, D., Rovera, G. (1987) Cytokine-dependent granulocytic differentiation. Regulation of proliferative and differentiative responses in a murine progenitor cell line. *J. Immunol.* 138, 3829–3835.
- Ymer, S., Tucker, W. Q., Sanderson, C. J., Hapel, A. J., Campbell, H. D., Young, I. G. (1985) Constitutive synthesis of interleukin-3 by leukaemia cell line WEHI-3B is due to retroviral insertion near the gene. *Nature* 317, 255–258.
- Collins, S. J., Ruscetti, F. W., Gallagher, R. E., Gallo, R. C. (1978) Terminal differentiation of human promyelocytic leukemia cells induced by dimethylsulfoxide and other polar compounds. *Proc. Natl. Acad. Sci. USA* 75, 2458–2462.
- Rovera, G., Santoli, D., Damsky, C. (1979) Human promyelocytic leukemia cells in culture differentiate into macrophage-like cells when treated with a phorbol diester. *Proc. Natl. Acad. Sci. USA* 76, 2779–2783.
- Paul, C. C., Aly, E., Lehman, J. A., Page, S. M., Gomez-Cambronero, J., Ackerman, S. J., Baumann, M. A. (2000) Human cell line that differentiates to all myeloid lineages and expresses neutrophil secondary granule genes. *Exp. Hematol.* 28, 1373–1380.
- Bhatia, M., Kirkland, J. B., Meckling-Gill, K. A. (1995) Modulation of poly(ADP-ribose) polymerase during neutrophilic and monocytic differentiation of promyelocytic (NB4) and myelocytic (HL-60) leukaemia cells. *Biochem. J.* 308, 131–137.
- Hubank, M., Schatz, D. G. (1994) Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Res.* 22, 5640–5648.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6.
- Nielsen, H., Krogh, A. (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6)*, Menlo Park, CA, AAAI, 122–130.
- Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016.
- Ouellette, B. F., Boguski, M. S. (1997) Database divisions and homology search files: a guide for the perplexed. *Genome Res.* 7, 952–955.
- Blake, J. A., Eppig, J. T., Richardson, J. E., Bult, C. J., Kadin, J. A. (2001) The mouse genome database (MGD): integration nexus for the laboratory mouse. *Nucleic Acids Res.* 29, 91–94.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., et al. (2001) The sequence of the human genome. *Science* 291, 1304–1351.
- Schuler, G. D. (1999) Sequence mapping by electronic PCR. *Genome Res.* 7, 541–550.
- Tatusova, T. A., Madden, T. L. (1999) Blast 2 sequences—a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250.
- Ohler, U., Niemann, H., Liao, G., Rubin, G. M. (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics* 17(Suppl 1) S199–206.
- Reese, M. G., Eeckman, F. H. (1995) Novel neural network algorithms for improved eukaryotic promoter site recognition. In *Proceedings of The Seventh International Genome Sequencing and Analysis Conference*, Hilton Head Island, SC, Hyatt Regency.
- Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Prüss, M., Reuter, I., Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28, 316–319.
- Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* 23, 4878–4884.
- Pustell, J., Kafatos, F. C. (1982) A high speed, high capacity homology matrix: zooming through SV40 and polyoma. *Nucleic Acids Res.* 10, 4765–4782.
- Thomas, P. S. (1983) Hybridization of denatured RNA transferred or dotted nitrocellulose paper. *Methods Enzymol.* 100, 255–266.
- Rovera, G., Santoli, D., Damsky, C. (1979) Human promyelocytic leukemia cells in culture differentiate into macrophage-like cells when treated with a phorbol diester. *Proc. Natl. Acad. Sci. USA* 76, 2779–2783.
- Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Bentley, D. R., et al. (1998) A physical map of 30,000 human genes. *Science* 282, 744–746.



30. Lekstrom-Himes, J., Xanthopoulos, K. G. (1998) Biological role of the CCAAT/enhancer-binding protein family of transcription factors. *J. Biol. Chem.* 273, 28545–28548.
31. Yamanaka, R., Lekstrom-Himes, J., Barlow, C., Wynshaw-Boris, A., Xanthopoulos, K. G. (1998) CCAAT/enhancer binding proteins are critical components of the transcriptional regulation of hematopoiesis (Review). *Int. J. Mol. Med.* 1, 213–221.
32. Zhang, D. E., Zhang, P., Wang, N. D., Hetherington, C. J., Darlington, G. J., Tenen, D. G. (1997) Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein alpha-deficient mice. *Proc. Natl. Acad. Sci. USA* 94, 569–574.
33. Yamanaka, R., Barlow, C., Lekstrom-Himes, J., Castilla, L. H., Liu, P. P., Eckhaus, M., Decker, T., Wynshaw-Boris, A., Xanthopoulos, K. G. (1997) Impaired granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-deficient mice. *Proc. Natl. Acad. Sci. USA* 94, 13187–13192.
34. Fisher, R. C., Scott, E. W. (1998) Role of PU.1 in hematopoiesis. *Stem Cells* 16, 25–37.
35. Simon, M. C. (1998) PU.1 and hematopoiesis: lessons learned from gene targeting experiments. *Semin. Immunol.* 10, 111–118.
36. Scott, E. W., Simon, M. C., Anastasi, J., Singh, H. (1994) Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science* 265, 1573–1577.
37. McKercher, S. R., Torbett, B. E., Anderson, K. L., Henkel, C. W., Vestal, D. J., Baribault, H., Klemsz, M., Feeney, A. J., Wu, G. E., Paige, C. J., Maki, R. A. (1996) Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO J.* 15, 5647–5658.
38. Ford, A. M., Bennett, C. A., Healy, L. E., Towatari, M., Greaves, M. F., Enver, T. (1996) Regulation of the myeloperoxidase enhancer binding proteins Pu1, C-EBP alpha, -beta, and -delta during granulocyte-lineage specification. *Proc. Natl. Acad. Sci. USA* 93, 10838–10843.
39. Oelgeschlaeger, M., Nuchprayoon, I., Luscher, B., Friedman, A. D. (1996) C/EBP, c-Myb, and PU.1 cooperate to regulate the neutrophil elastase promoter. *Mol. Cell. Biol.* 16, 4717–4725.
40. Smith, L. T., Hohaus, S., Gonzalez, D. A., Dziennis, S. E., Tenen, D. G. (1996) PU.1 (Spi-1) and C/EBP alpha regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. *Blood* 88, 1234–1247.
41. Bavisotto, L., Kaushansky, K., Lin, N., Hromas, R. (1991) Antisense oligonucleotides from the stage-specific myeloid zinc finger gene MZF-1 inhibit granulopoiesis in vitro. *J. Exp. Med.* 174, 1097–1101.
42. Ihle, J. N. (2001) The Stat family in cytokine signaling. *Curr. Opin. Cell Biol.* 13, 211–217.



# Gene Expression Profiling of Mucosal Addressin Cell Adhesion Molecule-1<sup>+</sup> High Endothelial Venule Cells (HEV) and Identification of a Leucine-Rich HEV Glycoprotein as a HEV Marker<sup>1</sup>

Koichi Saito,\* Toshiyuki Tanaka,\* Hidenobu Kanda,\* Yukihiro Ebisuno,\* Dai Izawa,\* Shoko Kawamoto,<sup>†</sup> Kosaku Okubo,<sup>†</sup> and Masayuki Miyasaka<sup>2\*</sup>

High endothelial venule (HEV) cells support lymphocyte migration from the peripheral blood into secondary lymphoid tissues. Using gene expression profiling of mucosal addressin cell adhesion molecule-1<sup>+</sup> mesenteric lymph node HEV cells by quantitative 3'-cDNA collection, we have identified a leucine-rich protein, named leucine-rich HEV glycoprotein (LRHG) that is selectively expressed in these cells. Northern blot analysis revealed that LRHG mRNA is ~1.3 kb and is expressed in lymph nodes, liver, and heart. In situ hybridization analysis demonstrated that the mRNA expression in lymph nodes is strictly restricted to the HEV cells, and immunofluorescence analysis with polyclonal Abs against LRHG indicated that the LRHG protein is localized mainly to HEV cells and possibly to some lymphoid cells surrounding the HEVs. LRHG cDNA encodes a 342-aa protein containing 8 tandem leucine-rich repeats of 24 aa each and has high homology to human leucine-rich  $\alpha_2$ -glycoprotein. Similar to some other leucine-rich repeat protein family members, LRHG can bind extracellular matrix proteins that are expressed on the basal lamina of HEVs, such as fibronectin, collagen IV, and laminin. In addition, LRHG binds TGF- $\beta$ . These results suggest that LRHG is likely to be multifunctional in that it may capture TGF- $\beta$  and/or other related humoral factors to modulate cell adhesion locally and may also be involved in the adhesion of HEV cells to the surrounding basal lamina. *The Journal of Immunology*, 2002, 168: 1050–1059.

Postnatally, circulating lymphocytes show a certain tissue tropism distinguishing venular endothelial cells in different sites of the body. This behavior is partly determined by the expression by the lymphocytes of adhesion molecules, such as L-selectin and  $\alpha_4\beta_7$  integrin, that can specifically recognize endothelial adhesion receptors expressed by venular endothelial cells in a tissue-specific manner (reviewed in Refs. 1–3). These vascular counterreceptors for lymphocyte adhesion molecules are termed vascular addressins, because they provide geographical cues or address codes to circulating lymphocytes (1, 4), and those expressed on high endothelial venule (HEV)<sup>3</sup> cells in lymph nodes (LNs) and Peyer's patches have been the most extensively studied.

These vascular addressins include the peripheral lymph node addressins (PNAd), which consist of core proteins that display sulfated mucin-type carbohydrates, such as GlyCAM-1 (5), CD34 (6), and podocalyxin (7), and the mucosal addressin, mucosal addressin cell adhesion molecule-1 (MAdCAM-1) (8). It is generally believed that PNAd interacts with L-selectin, directing lymphocytes to the peripheral LNs (5), whereas MAdCAM-1 interacts with  $\alpha_4\beta_7$  integrin, directing lymphocytes to mesenteric LNs and Peyer's patches (9).

HEV cells express not only vascular addressins but also certain chemokines that enable their specific interaction with lymphocytes. In particular, a CC chemokine, secondary lymphoid tissue chemokine (SLC) (10), is produced by HEV cells and can rapidly activate the  $\alpha_L\beta_2$  (LFA-1) (11) and  $\alpha_4\beta_7$  (12) integrins on lymphocytes, allowing the lymphocytes to adhere firmly to and trans-migrate across the HEV. *plt/plt* mice, which are genetically deficient in SLC expression by HEV, show extremely impaired migration of T cells and dendritic cells into the LNs and spleen (13), indicating the prime importance of this molecule in the trafficking of certain leukocyte subsets into lymphoid tissues. HEV cells also express a variety of other chemokines, including EBI-1 ligand chemokine (14) and B lymphocyte chemoattractant (15), but their functional significance remains to be clarified.

Given the redundancy observed in chemokines and adhesion molecules (16), it is not difficult to imagine that HEV cells express many more chemokines and adhesion molecules than we know of presently. To address this possibility, intensive investigation has been performed to identify novel molecules expressed specifically in HEVs, and a number of such molecules have been found, including hevin (17), HEV-specific *N*-acetylglucosamine 6-sulfo-transferase (18, 19), junctional adhesion molecule (JAM)-2 (20), vascular endothelial JAM (21), and endoglycan (22). In addition, although not novel, other molecules that were originally reported

\*Laboratory of Molecular and Cellular Recognition, Osaka University Graduate School of Medicine; and <sup>†</sup>Institute for Molecular and Cellular Biology, Osaka University, Suita, Japan

Received for publication April 10, 2001. Accepted for publication November 30, 2001.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> This work was supported in part by a Grant-in-Aid for COE Research from the Ministry of Education, Science, Sports and Culture, Japan; a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture, Japan; and a grant from the Science of Technology Agency, Japan.

<sup>2</sup> Address correspondence and reprint requests to Dr. Masayuki Miyasaka, Laboratory of Molecular and Cellular Recognition, Osaka University Graduate School of Medicine C8, 2-2, Yamada-oka, Suita, 565-0871, Japan. E-mail address: mmiasak@orgcl.med.osaka-u.ac.jp

<sup>3</sup> Abbreviations used in this paper: HEV, high endothelial venule; ApoE, apolipoprotein E; ECM, extracellular matrix; GlyCAM-1, glycosylation-dependent cell adhesion molecule-1; MAdCAM-1, mucosal addressin cell adhesion molecule-1; LN, lymph node; LRHG, leucine-rich HEV glycoprotein; LRR, leucine-rich repeat; PNAd, peripheral node addressin; SLC, secondary lymphoid tissue chemokine; JAM, junctional adhesion molecule; DARC, Duffy Ag/receptor for chemokine; rhTGF $\beta$ R, recombinant human TGF- $\beta$ RII/Fc chimera; GS, gene signature; MyD118, myeloid differentiation 118; SPI3, serine proteinase inhibitor 3; VE, vascular endothelial.

to be present in nonlymphoid tissues have been found in LNs, particularly in HEV cells. These include mac25/angiomodulin/IGFBP-rP1 (23, 24) and a promiscuous chemokine receptor, Duffy Ag/receptor for chemokine (DARC) (23, 24).

To identify novel molecules in HEV, we previously performed an unbiased gene expression analysis in mouse HEV cells obtained from peripheral LNs (23). In our previous analysis, we prepared a 3'-directed cDNA library that faithfully represented the original mRNA composition of highly purified MECA-79 (PNAd)-positive mouse HEV cells and analyzed ~1500 3'-cDNA sequences randomly selected from the library. Subsequently, by comparing these sequences with those obtained from 35 cell types, we found that MECA-79<sup>+</sup> peripheral LN HEV cells exhibit a unique gene expression profile that includes a few novel genes (23).

In the present study, we extended this analysis to MECA-367 (MAdCAM-1)-positive HEV cells and herein report a list of genes selectively expressed in mouse mesenteric HEV cells. The expression profiling analysis also allowed us to identify as a novel HEV marker a 342-aa protein that contains tandem arrays of the leucine-rich repeat (LRR) motif. This protein apparently belongs to the LRR superfamily (25) and is likely to be a mouse homolog of leucine-rich  $\alpha_2$ -glycoprotein, a protein previously identified in human plasma (26). Because the mRNA and protein product are abundantly and selectively expressed in HEV, we designated this molecule leucine-rich HEV glycoprotein (LRHG). LRHG interacted with various extracellular matrix (ECM) proteins, similar to other LRR proteins, and also bound TGF- $\beta$ . These findings suggest that LRHG may be involved in the regulation of HEV-ECM interactions as well as in modulating the adhesive properties of lymphoid cells.

## Materials and Methods

### Animals and reagents

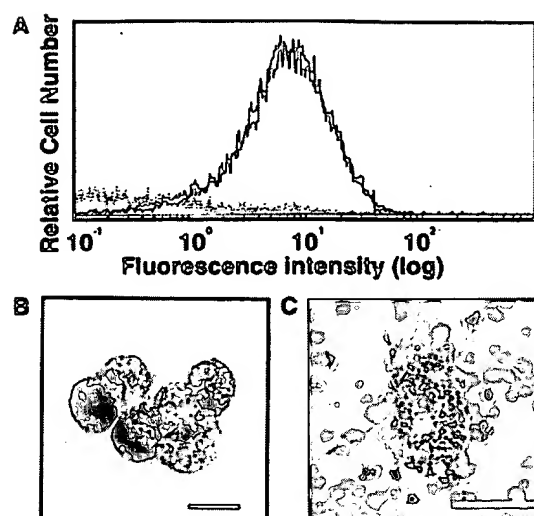
All animal experiments were performed under the experimental protocol approved by the Ethics Review Committee for Animal Experimentation of Osaka University Graduate School of Medicine. Male C57BL/6 mice were purchased from Japan SLC (Hamamatsu, Japan). MECA-79 (27) and MECA-367 mAbs (4) were kindly provided by Dr. E. C. Butcher (Stanford University, Stanford, CA). The following ECM proteins were obtained from commercial sources: mouse type IV collagen, natural mouse laminin, mouse fibronectin, human type I and type III collagen, and human vitronectin were purchased from Life Technologies (Gaithersburg, MD); human type V and type VI collagen were from Southern Biotechnology (Birmingham, AL); mouse type II collagen was from Elastin Products (Owensville, MO); human fibronectin was from ICN Pharmaceuticals (Aurora, OH); recombinant human TGF- $\beta$ RII/Fc chimera (rhTGF $\beta$ R) and monoclonal anti-human TGF $\beta$ 1 Ab (anti-TGF- $\beta$  mAb) were purchased from R&D Systems (Minneapolis, MN).

### Construction of cDNA library and sequencing

HEV cells were isolated from mouse mesenteric LNs using the MECA-367 mAb by immunomagnetic selection, and total RNA was prepared from purified MECA-367<sup>+</sup> cells as described by Izawa et al. (23). Using 160 ng total RNA from purified MECA-367<sup>+</sup> cells, a 3'-directed cDNA library was constructed and analyzed as described previously (23).

### Cell culture

The mouse LN-derived endothelial cell lines KOP2.16 (28), HEC367-1, and HEC367-2 were maintained in DMEM (Sigma-Aldrich, St. Louis, MO) supplemented with 20% heat-inactivated FCS (HyClone Laboratories, Logan, UT), 10 mM HEPES, 1 mM sodium pyruvate, 2 mM L-glutamine, 1% (v/v) 100 $\times$  nonessential amino acids, 100 U/ml penicillin, and 100  $\mu$ g/ml streptomycin. The mouse endothelial cell lines, F-2 (29) and SVEC4-10 (30), and a mouse fibroblast L cell line were maintained in DMEM containing 10% heat-inactivated FCS (Dainippon Pharmaceutical, Osaka, Japan) and the same supplements described above.



**FIGURE 1.** Purification of MECA-367<sup>+</sup> HEV cells from mouse mesenteric LN. *A*, Flow cytometric analysis of purified MECA-367<sup>+</sup> HEV cells (—) and the stromal cell fraction (.....). *B*, Purified MECA-367<sup>+</sup> HEV cells were spun in a cytocentrifuge and stained with the MECA-367 mAb. Black bar, 10  $\mu$ m. *C*, Lymphocyte binding to purified MECA-367<sup>+</sup> HEV cells. Purified MECA-367<sup>+</sup> HEV cells were cultured, and added lymphocytes avidly transmigrated underneath them. White bar, 100  $\mu$ m.

### Lymphocyte-endothelial cell adhesion assay

Purified MECA-367<sup>+</sup> cells were plated at semiconfluent density in six-well plates. LN cells were then added to the wells and incubated for 2 h at 37°C. Nonadherent lymphocytes were removed by gentle washing, and then lymphocyte adhesion to and transmigration underneath the endothelial cells were observed by microscopy.

### Northern blot analysis

Mouse poly(A)<sup>+</sup> multitissue Northern blots (Clontech, Palo Alto, CA) were hybridized with a <sup>32</sup>P-labeled LRHG or  $\beta$ -actin probe (1.0  $\times$  10<sup>6</sup> cpm/ml) using ExpressHyb hybridization buffer (Clontech). RNA from various mouse tissues was isolated using TRIzol (Life Technologies) according to the manufacturer's instructions. The samples were fractionated on a 0.8% agarose gel containing 17% formaldehyde and transferred to Hybond-N<sup>+</sup> nylon membranes (Amersham Pharmacia Biotech, Uppsala,

**Table 1.** Gene expression profile of MECA-367<sup>+</sup> HEV cells<sup>a</sup>

	No. of GS	No. of Sequences
Known genes	427	1016
Genes with housekeeping functions	118	405
Ribosomal proteins	58	307
Protein synthesis and degradation	19	31
Nuclear proteins	14	19
Vesicle trafficking and transporters	15	26
Energy production	12	22
Genes with specialized functions	309	611
Cytoskeleton-related proteins	29	60
Cytokines/ECM proteins	22	47
Plasma membrane proteins	42	110
Intracellular signaling	53	81
Transcription factors	18	25
Proteinases and proteinase inhibitors	11	23
Cellular enzymes	52	64
Other functions	82	201
Novel genes	877	1085
<b>Total</b>	<b>1304</b>	<b>2101</b>

<sup>a</sup> The sequences of 2101 independent clones from a 3'-directed cDNA library were grouped into 1304 GS species according to sequence identity. These GS species were then divided into those that matched known genes and those that did not. The former group was further categorized according to function and is listed here.

Table II. Comparative analysis of gene expression in HEV cells<sup>a</sup>

Function/Definition	Recurrence (total no. of sequences)						GenBank
	M/HE (2101)	P/HE (1558)	FE (1671)	T4 (1128)	T8 (1446)	B (1059)	
Cytoskeleton related proteins							
Thymosin $\beta$ -4	13	25	20	10	11	5	X16053
Cytoskeletal $\gamma$ -actin	5	3	5	1	1	0	M21495
Myosin L chain 3	3	3	6	0	0	2	U04443
Myelin-regulatory factor 1	3	2	0	0	0	0	U14648
Cytoplasmic $\gamma$ -actin	3	1	5	0	0	0	L21996
$\beta$ -Actin	3	0	1	2	2	2	X03672
$\beta$ -Tubulin	3	0	1	0	0	0	X03369
Myosin-regulatory light chain	3	0	4	1	0	1	X05566
Cytokines and ECM proteins							
SLC <sup>b</sup>	12	8	0	0	0	0	AF001980
Follistatin-like protein mac25	11	9	0	0	0	0	L75822
TAG7 protein	6	6	0	0	0	0	X86374
<i>fau</i>	5	4	3	3	6	5	X65922
SPARC	3	1	5	0	0	0	X04017
KC protein	2	2	0	0	0	0	J04596
Epithelin 1 and 2/acrogranin	2	1	0	0	0	0	X62321
SDF2	2	0	0	0	0	0	D50646
$\alpha_1$ -chain of collagen type IV	2	0	0	0	0	0	X92439
GlyCAM-1 <sup>c</sup>	0	2	0	0	0	0	M93428
Proteinase/inhibitors							
<i>ctla-2<math>\alpha</math></i>	6	6	8	1	0	0	X15591
$\alpha_1$ -PI-1	4	6	0	0	0	0	M75721
Cathepsin L	3	5	1	0	0	0	M20495
SPI3	3	0	0	0	0	0	U25844
Plasma membrane proteins							
MAdCAM-1 <sup>d</sup>	11	0	0	0	0	0	L21203
$\beta_2$ m	8	3	1	1	2	2	J00365
Chemokine receptor DARC	7	3	0	0	0	0	AF016697
Thymic shared Ag-1	6	4	0	0	0	4	U47737
L6 Ag	6	2	8	0	0	0	L15429
MHC class I H2-K	6	1	0	4	1	2	J00400
IL-3R	6	1	0	0	0	0	M29855
VE-cadherin	6	0	0	0	0	0	X83678
Endoglin	5	7	1	0	0	0	X77952
Lactadherin	5	0	0	0	0	0	M38337
Ly-6C.2	3	0	0	0	0	0	M18466
IL-2R $\gamma$ -chain	3	0	0	0	0	0	D13565
ICAM-1	2	1	1	0	0	2	M90551
JAM	2	0	0	0	0	0	U89915
Intracellular signaling							
Phosphodiesterase I	5	5	0	0	0	0	D28560
CDC42	4	0	0	1	2	1	U37720
G protein $\beta$	4	0	3	4	3	4	D29802
<i>bcl-3</i>	3	0	0	0	0	1	M90397
Cyclin D1	3	0	0	0	0	0	S78355
G- $\alpha$ -i2	3	0	0	0	1	0	M17528
Calmodulin	3	0	0	1	0	0	M19381
ADP ribosylation factor	3	2	0	0	0	0	M86705
<i>ras</i> -related protein p32	2	2	0	0	0	0	X12535
Cab45a	2	2	0	0	0	0	U45977
<i>erp</i>	2	2	0	0	0	0	S64851
Adenylyl cyclase type IV	2	1	0	0	0	0	M80633
SH3P2	2	0	0	0	0	0	U58888
<i>rab11</i>	2	0	0	0	0	0	D50500
Transcription factors							
LRG-21	4	2	0	0	0	0	U19118
CCR4 protein	2	2	0	0	0	0	U70139
Y-box-binding protein 1	2	2	2	3	1	1	M60419
<i>Ets-1</i>	2	0	0	0	0	0	M31885

<sup>a</sup> Genes appearing more than three times and some of those twice in the MECA-367<sup>+</sup> HEV library are listed according to their function or cellular localization. Numbers represent the frequency among the 1304 GS analyzed. The abundance of the GS in other libraries (see <http://bodymap.ims.u-tokyo.ac.jp/> for details) is also shown for comparison. M/HE, MECA-367<sup>+</sup> HEV cells analyzed in the present study; P/HE, MECA-79<sup>+</sup> HEV cells analyzed in our previous study (23); FE, CD31<sup>+</sup> flat endothelial cells; T4, CD4<sup>+</sup> T cells; T8, CD8<sup>+</sup> T cells; B, B220<sup>+</sup> B cells; SPARC, secreted protein acidic and rich in cysteine; SDF-2, stromal cell-derived factor-2;  $\alpha_1$ -PI-1,  $\alpha_1$  protease inhibitor 1; G- $\alpha$ -i2, GTP-binding protein- $\alpha$ -i2; NDP, nucleoside diphosphate; SH3P2, SH3 containing protein 2.

<sup>b</sup> Transcripts for SLC appeared as a 3'-end cDNA sequence of 14 bp which had been excluded from analysis in the previous study (23).

<sup>c</sup> Transcripts for GlyCAM-1 abundantly expressed in MECA-79<sup>+</sup> HEV cells appeared as aberrantly digested forms (23).

<sup>d</sup> Transcripts for MAdCAM-1 terminating with different poly(A) sites are regarded as identical in this table.

Table II. *continued*

Function/Definition	Recurrence (Total no. of sequences)						GenBank
	M/HE (2101)	P/HE (1558)	FE (1671)	T4 (1128)	T8 (1446)	B (1059)	
Cellular enzymes							
Transglutaminase	6	8	0	0	0	0	M55154
$\alpha$ -Amylase-2	3	0	0	0	0	0	J00360
NDP kinase B	2	1	4	1	0	2	X68193
3,2-trans-Enoyl-CoA isomerase	2	0	0	0	0	1	Z14049
Ornithine decarboxylase	2	0	0	0	0	0	M10624
GST II	2	0	2	0	1	1	X53451
Other function							
ApoE	15	12	0	0	0	0	M12414
24.6-kDa protein	15	9	14	8	11	6	M93980
21-kDa polypeptide	12	0	10	19	20	7	X06407
IFN-induced mRNA	11	11	1	0	0	0	X61381
tum-Ag p198	9	3	8	28	29	10	X51528
Clusterin/ApoJ/sgp-2	8	2	1	0	0	0	D14077
Insulinoma (rig)	7	9	13	4	15	9	M33330
Cyclophilin	7	1	16	2	5	8	X52803
Ferritin H chain	5	2	2	1	1	2	X12812
Ferritin L chain	5	2	7	3	2	1	J04716
<i>Mus musculus</i> mRNA	5	1	0	0	0	0	L29441
Laminin receptor	5	1	8	3	11	5	J02870
gly96	5	0	1	0	0	0	X67644
Growth factor-induced protein	3	6	0	0	0	0	L02914
FKBP-12 pseudogene-1	3	1	2	0	0	0	U65094
Glucose-regulated protein 78	3	1	0	1	0	0	M30779
MyD118	3	0	0	0	0	0	X5414

Sweden). The filters were hybridized with the LRHG or  $\beta$ -actin probe as described above. In certain tissues, such as skeletal muscle, kidney, and testis, the  $\beta$ -actin probe detected not only the standard 2-kb band but also smaller bands (1.6–1.8 kb), which represent  $\beta$ -actin isoforms.

#### cDNA library screening and cloning of LRHG cDNA

A cDNA fragment of LRHG was labeled with HRP using the ECL direct nucleic acid labeling and detection system (Amersham) and was used to screen a mouse liver 5'-STRECH PLUS Triplex cDNA library (Clontech). Approximately  $1.2 \times 10^5$  PFU were immobilized on Hybond-N<sup>+</sup> nylon membrane and then hybridized with the HRP-labeled probe (700 pg/ml) according to the manufacturer's instructions.

#### RT-PCR analysis

First-strand cDNA synthesis from total RNA (1  $\mu$ g) was performed using Ready-To-Go (Amersham) with an oligo(dT) primer. PCR was conducted using a sense primer (5'-GATGATGGCTGGGGTGTGCTG-3') and an antisense primer (5'-AACTGCTTTGGTGACCCCTGAAAC-3') specific to mouse LRHG and ExTaq polymerase (TaKaRa, Otsu, Japan) under the following conditions: 94°C for 1 min; 94°C for 30 s, 60°C for 30 s, 72°C for 1 min, 27 cycles; 72°C for 5 min. As a control, a primer pair for mouse  $\beta$ -actin (5'-ATGGATGACGATATCGC-3' and 5'-ATGAGGTAGTCTGTCAGGT-3') was used. PCR products were analyzed by agarose gel electrophoresis.

#### In situ hybridization

The LRHG cDNA in pTriplex plasmid was transcribed into digoxigenin-labeled antisense RNA with T3 polymerase (Stratagene, La Jolla, CA) or sense RNA with T7 polymerase (Toyobo, Osaka, Japan), using the DIG RNA Labeling Mix (Boehringer Mannheim, Mannheim, Germany). Frozen mesenteric LN sections (10  $\mu$ m) from C57BL/6 mice were hybridized with digoxigenin-labeled RNA probe (10 ng/ $\mu$ l) and reacted with 1.5 U/ml alkaline phosphatase-conjugated anti-digoxigenin (Boehringer Mannheim).

#### Production of recombinant LRHG and its purification

An open reading frame of LRHG cDNA was inserted into a pBAD-myc-His vector (Invitrogen, Carlsbad, CA). The resulting expression plasmid containing LRHG-myc-His was transfected into Top10 *Escherichia coli* (Invitrogen). The recombinant protein was purified using the Xpress System (Invitrogen). Briefly, the *E. coli* cell lysates were sonicated and sub-

jected to rapid freeze-thaw cycles. The rLRHG protein was purified from the cleared lysate using Ni<sup>2+</sup>-charged columns following the manufacturer's recommendations. This rLRHG was used in the ECM-binding assay.

To isolate inclusion bodies from LRHG-transformants, the *E. coli* were suspended in 10 mM KH<sub>2</sub>PO<sub>4</sub> (pH 7.0) and 1 mM EDTA (lysis buffer) and sonicated. After centrifugation, the pellet was resuspended in lysis buffer containing 0.5% Triton X-100. Insoluble materials were then washed with H<sub>2</sub>O and solubilized in 8 M urea. The rLRHG was purified using a Ni<sup>2+</sup>-charged agarose column and also an anti-myc mAb-conjugated agarose column (Santa Cruz Biotechnology, Santa Cruz, CA). This rLRHG preparation was used in the TGF- $\beta$ -binding assay.

#### Generation of rabbit polyclonal Abs against mouse LRHG

Polyclonal Abs were raised against rLRHG protein by s.c. immunization of rabbits with the protein (100  $\mu$ g), which had been emulsified in TiterMax Gold (CytRx, Norcross, GA). The polyclonal IgG was affinity purified from immunized rabbit serum using a protein G (Amersham) column and a column conjugated with a LRHG peptide (PADTVHLSVEFS, corresponding to aa 60–71).

#### Immunohistochemistry

Immunostaining of frozen sections was performed as previously described (31). Briefly, mesenteric LN cryosections that were fixed in acetone and then in 4% paraformaldehyde in PBS were incubated with the anti-LRHG polyclonal Ab. The sections were then incubated with biotin-conjugated anti-rabbit IgG, followed by alkaline phosphatase-conjugated ABC reagent (Vector Laboratories, Burlingame, CA). After gentle fixation in 1% glutaraldehyde, the sections were stained using Vector Red (Vector Laboratories) as a substrate. For two-color staining, the sections were further incubated with FITC-conjugated MECA-367 or MECA-79 mAbs. Purified MECA-367<sup>+</sup> cells were spun in a cytocentrifuge, fixed with methanol, and incubated with biotin-conjugated MECA-367 mAb. After a washing, the cells were stained with ABC reagent (Vector) and Metal Enhanced DAB (Pierce, Rockford, IL).

#### LRHG binding to ECM proteins

Various ECM proteins (10  $\mu$ g/ml) dissolved in 0.1 M Tris-HCl (pH 7.4), 50 mM NaCl were immobilized onto 96-well microtiter plates (Sumilon H; Sumitomo Bakelite, Tokyo, Japan) at 4°C (50  $\mu$ l/well) (32). The wells were blocked with 3% BSA and incubated with myc-His-tagged LRHG or myc-His-tagged mac25 (provided by D. Nagakubo of our laboratory) or

myc-His-tagged L-selectin (provided by H. Kawashima of our laboratory; 10  $\mu$ g/ml). Binding of the rLRHG was detected with HRP-conjugated anti-myc mAb (Invitrogen) and *o*-phenylenediamine as a substrate.

#### *Binding of $^{125}$ I-TGF- $\beta$ to recombinant LRHG*

Recombinant myc-His-tagged LRHG, myc-His-tagged L-selectin, rhTGF $\beta$ R, and anti-TGF- $\beta$  mAb dissolved in PBS (10  $\mu$ g/ml) were immobilized onto 96-well microtiter plates (50  $\mu$ l/well) at 4°C. After blocking with PBS containing 1% BSA and 0.05% Tween 20 (TPBS),  $^{125}$ I-TGF- $\beta$  (Amersham) was added to each well, and the plates were incubated for 4 h at 37°C. The wells were then washed with TPBS, and the bound radioactivity was counted. To verify binding specificity, rLRHG was first incubated with anti-myc mAb-conjugated agarose (Santa Cruz Biotechnology) or Ni $^{2+}$ -charged agarose (Invitrogen). Unbound materials were immobilized onto 96-well microtiter plates, and the binding assay was performed as described above.

## Results

### *Purification of MECA-367 $^{+}$ HEV cells from mouse mesenteric LN*

We first obtained an HEV-enriched stromal cell fraction from mouse mesenteric LNs, as we described previously (23). This stromal cell fraction was substantially enriched with HEV cells; 7.5% of the cells were HEV cells as assessed by immunofluorescence staining with the MAdCAM-1-specific mAb, MECA-367 (normally, <0.01% of the LN cells are MAdCAM-1 $^{+}$ ). After gentle trypsin digestion, the stromal cell fraction was further subjected to two rounds of immunomagnetic cell sorting with MACS using the MECA-367 mAb, and, as assessed by flow cytometry, over 90% of the sorted cells were MECA-367 $^{+}$  (Fig. 1A). Immunoperoxidase staining of the sorted cells showed that the majority of cells were large nonlymphoid cells expressing MAdCAM-1 (Fig. 1B). In addition, when these cells were plated on culture dishes, mesenteric LN lymphocytes bound to and migrated underneath them avidly, indicating that they have the phenotype and function consistent with their identification as HEV cells (Fig. 1C).

### *Analysis of 3'-directed cDNA library from mouse MECA-367 $^{+}$ HEV cells*

To investigate the gene expression profile of HEV cells, we constructed a 3'-directed cDNA library from the purified MECA-367 $^{+}$  HEV cells as we had previously done with MECA-79 $^{+}$  HEV cells (23) and subjected these cDNAs to cycle sequencing reactions to collect a total of 2101 cDNA sequences. These short 3'-cDNA sequences are called gene signatures (GSs), because they are unique to individual genes (33). The GS sequences with >90% identity were regarded as identical. Accordingly, they were grouped together and subsequently classified into 1304 independent GSs. By examining these sequences against the GenBank database, we found that 427 of the GSs were derived from genes that had been previously reported, and 877 of them were from unidentified genes. Among the known genes, about one-third encoded ribosomal proteins (Table I), consistent with the notion that HEV cells have intense biosynthetic activity (34).

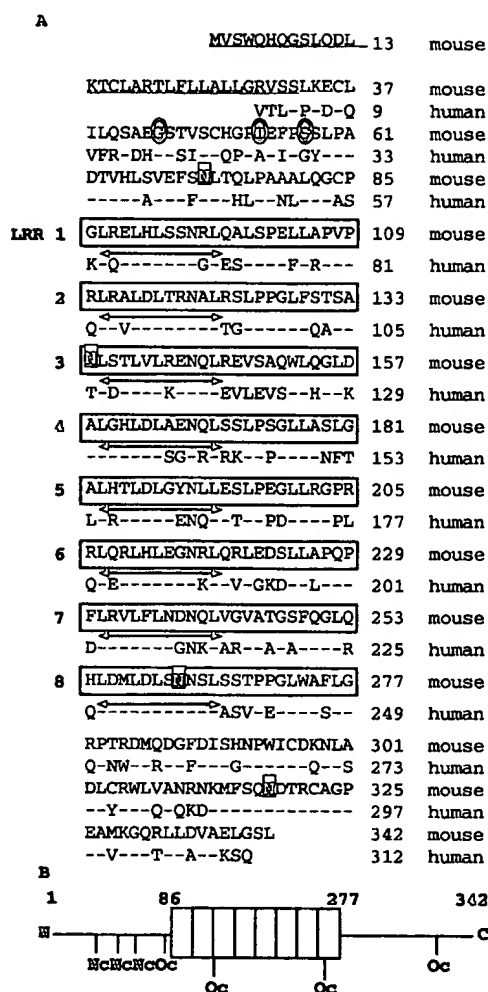
We next compared the gene expression profile of mouse MECA-367 $^{+}$  HEV cells with expression profiles obtained from other cell types, including MECA-79 $^{+}$  HEV cells, CD31 $^{+}$  flat endothelial cells, and T and B lymphocytes. Table II lists the differentially expressed genes that are already known. Several important points can be drawn from this analysis. First, the tissue-specific vascular addressins, MAdCAM-1 and GlyCAM-1, are expressed selectively in MECA-367 $^{+}$  HEV cells and MECA-79 $^{+}$  HEV cells, respectively, but not in other cell types, validating the previous histological observations (4, 27) and the HEV cell separation protocol used in the present study. Second, as observed in CD31 $^{+}$  flat endothelial cells and MECA-79 $^{+}$  HEV cells, a number of endothelial cell markers such as endoglin (35), ICAM-1, and L6 Ag are also expressed in MECA-367 $^{+}$  HEV cells, confirming that

the MECA-367 $^{+}$  HEV cells bear endothelial phenotypes. Third, a few molecules are preferentially expressed in MECA-79 $^{+}$  HEV cells and MECA-367 $^{+}$  HEV cells, but not in CD31 $^{+}$  flat endothelial cells. These include certain chemokines, such as SLC (10) and KC (36), an inflammatory cytokine, TAG7 (37), a promiscuous chemokine receptor, DARC (38), and some TGF- $\beta$ -responsive molecules, like mac25 (39), transglutaminase (40), and apolipoprotein E (ApoE) (41). Because these molecules are apparently not expressed in CD31 $^{+}$  flat endothelial cells, they may exert unique functions in HEV cells, possibly playing a role in conferring some of the unique properties that these endothelial cells possess. Fourth, certain transcripts may be preferentially expressed in MECA-367 $^{+}$  cells and not in MECA-79 $^{+}$  cells. However, the recurrence of these transcripts in the present analysis was only low to moderate, so their significance remains unclear. These include vascular endothelial (VE)-cadherin, lactadherin, Ly-6C.2, IL-2R  $\gamma$ -chain, serine proteinase inhibitor 3 (SPI3), cyclin D1,  $\alpha$ -amylase-2, and myeloid differentiation 118 (MyD118).

### *Identification of a novel member of the LRR protein family in MECA-367 $^{+}$ HEV cells*

Further comparison of the gene expression profile of the MECA-367 $^{+}$  HEV-derived cDNA library with expression profiles obtained from 35 tissues and cell types and the GenBank database revealed several hitherto unidentified genes in the mouse to be highly expressed in MECA-367 $^{+}$  HEV cells. Because one of them (GS12070) was also highly expressed in the liver (see below), we isolated a full length cDNA from a mouse liver cDNA library and determined its complete nucleotide sequence. The full length cDNA was 1.3 kb long and contained a single open reading frame that began at nt 21 and terminated at nt 1049, encoding a putative protein of 342 aa (Fig. 2). Because a cDNA obtained from the MECA-367 $^{+}$  HEV cells had an identical nucleotide sequence (data not shown), we reasoned that the same protein was expressed in the liver and MECA-367 $^{+}$  HEV cells. The deduced amino acid sequence contained tandem arrays of 8 LRRs; this motif is found in >60 proteins and is thought to be involved in protein-protein interactions (25). As seen in other LRR protein family members, each LRR of this protein contained a well-conserved 11-residue segment (LxxLxLxxN/CxL). This LRR protein had four potential N-linked glycosylation sites, three potential O-linked glycosylation sites, and no apparent transmembrane domain. Although some LRR proteins are proteoglycans, e.g., decorin and biglycan, this LRR protein had no potential glycosaminoglycan attachment sites, such as serine-glycine and serine-alanine pairs. It shared an extremely similar domain structure and high amino acid homology (67%) with human leucine-rich  $\alpha_2$ -glycoprotein, a protein of unknown function initially identified in human plasma (26), and hence was likely to be its mouse homolog. Because the high expression of this unique protein in HEV has not been reported before, we designated this protein LRHG.

We then performed Northern blotting analysis using a fragment of the LRHG cDNA as a probe. As shown in Fig. 3, a major 1.3-kb band was detected in the liver, heart (Fig. 3A), and lymphoid tissues (Fig. 3B), and a 2.8-kb band, presumably representing an alternatively spliced form, was also detected in the liver (Fig. 3A). Among the lymphoid tissues, this mRNA (the 1.3-kb band) was expressed in the LNs but not spleen or thymus. In situ hybridization analysis verified that LRHG mRNA is highly expressed in mesenteric HEV cells, similar to MAdCAM-1 mRNA and GlyCAM-1 mRNA, whereas no signal was obtained with a sense probe for any of these glycoproteins (Fig. 4). In addition, RT-PCR analysis indicated that purified MECA-79 $^{+}$  cells and MECA-367 $^{+}$  cells, but no other endothelial cell lines examined, expressed



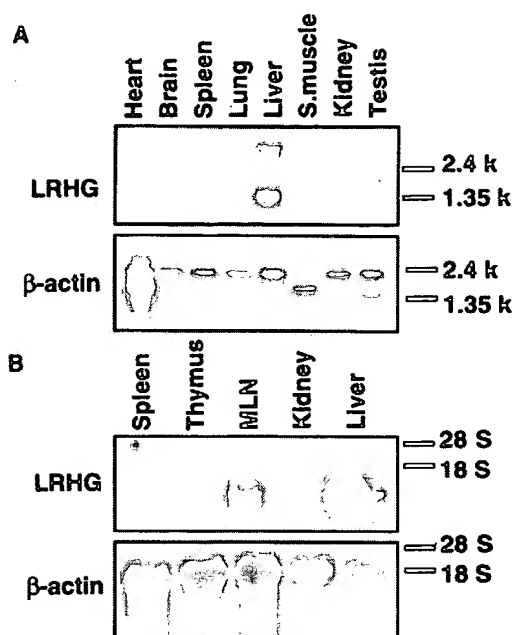
**FIGURE 2.** Deduced LRHG amino acid sequence. **A**, Alignment of deduced amino acid sequence of LRHG and human leucine-rich  $\alpha_2$ -glycoprotein. The identical amino acid residues are shown by a bar. The predicted signal peptide sequence is underlined. Putative N-glycosylation sites are in black boxes, and putative O-glycosylation sites are in circles. The LRR domains are in boxes and conserved  $\beta$  sheets in the LRR domains are indicated by arrows. **B**, Schematic illustration of the LRHG protein. The boxes represent the LRR domains. LRHG has eight LRRs. The numbers indicate the positions of amino acid residues. Nc, putative N-glycosylation site; Oc, putative O-glycosylation site. The nucleotide sequence has been submitted to the EMBL Data Library/GenBank/DBJ databases with the accession number AB055885.

LRHG mRNA (Fig. 5). This finding suggested that both types of HEV cells commonly express LRHG and may lose their expression during in vitro culture.

Immunohistological staining of frozen sections of mesenteric LN showed that LRHG was localized to HEV cells and the areas surrounding the HEVs (Fig. 6). Because LRHG mRNA is expressed only in HEV cells (Fig. 4), this observation may indicate that LRHG is secreted by HEV cells and sequestered in their vicinity.

#### LRHG binds to extracellular matrix

Because some of the LRR proteins that are structurally related to LRHG, such as decorin and biglycan, bind various ECM proteins (42), we next sought to determine whether LRHG could also bind ECM proteins. For this purpose, myc-His-tagged LRHG was added to wells containing immobilized ECM proteins, and after a washing, LRHG binding was examined using an anti-myc mAb.



**FIGURE 3.** Northern blot analysis of LRHG distribution in mouse tissues. **A**, A mouse tissue poly(A) RNA blot was probed with  $^{32}$ P-labeled LRHG or  $\beta$ -actin cDNA. The sizes of the RNA standards are indicated in kilobases (k). See *Materials and Methods* for an explanation of the multiple  $\beta$ -actin bands seen in certain tissues. **B**, The total RNA (30  $\mu$ g) from various mouse tissues was separated in a 0.8% agarose gel and hybridized with a  $^{32}$ P-labeled LRHG or  $\beta$ -actin probe. The molecular size of ribosomal RNA (28S and 18S) is indicated. MLN, mesenteric LN.

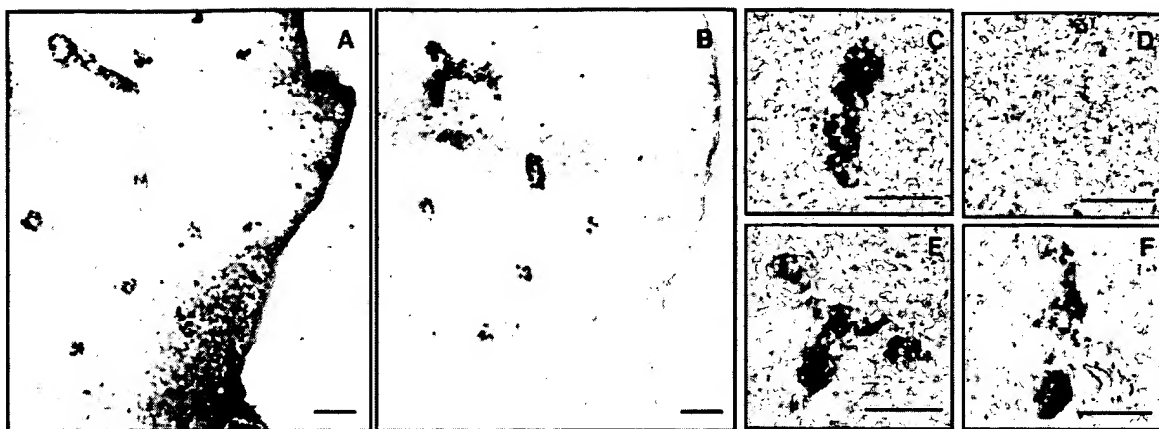
As shown in Fig. 7, LRHG bound to fibronectin, laminin, and various types of collagens moderately, whereas another HEV protein, mac25, bound to type IV collagen strongly (39). A control protein L-selectin that was also myc-His-tagged, similar to LRHG and mac25, did not bind to any of the ECM proteins examined. These results demonstrate that LRHG can bind various ECM proteins, particularly those that accumulate in the basal lamina of vascular beds, and suggest that LRHG may participate in regulating the adhesive interactions of HEV cells with the surrounding ECM proteins in the basal lamina.

#### LRHG binds TGF- $\beta$

Certain LRR proteins have been reported to bind TGF- $\beta$  (43). Therefore, we sought to determine whether LRHG can also bind TGF- $\beta$ . Recombinant myc-His-tagged LRHG was immobilized onto a plastic support and subjected to a binding assay with  $^{125}$ I-labeled TGF- $\beta$ . As shown in Fig. 8A, myc-His-tagged LRHG protein bound TGF- $\beta$ , whereas this binding was not as strong as that of rhTGF- $\beta$ R or anti-TGF- $\beta$  mAb that was used as a positive control in this experiment. As shown in Fig. 8B, LRHG binding to TGF- $\beta$  was specific and not mediated by a minor contaminant(s) in the LRHG preparation we used, because absorption of the recombinant protein with anti-myc-conjugated beads or a  $\text{Ni}^{2+}$ -charged column abrogated the TGF- $\beta$  binding. Another LRR protein, decorin, also bound TGF- $\beta$  (data not shown) as previously demonstrated (43).

#### Discussion

In the present study, we purified mouse mesenteric MAdCAM-1 $^{+}$  HEV cells, constructed a 3'-directed cDNA library, and obtained a gene expression profile of MAdCAM-1 $^{+}$  HEV cells by the single-cycle sequencing of 2101 clones randomly selected from the



**FIGURE 4.** In situ hybridization analysis of the LRHG in mouse mesenteric LN. Sections were hybridized with a digoxigenin-labeled LRHG antisense riboprobe (A), MADCAM-1 antisense riboprobe (B), LRHG antisense riboprobe (C), LRHG sense probe (D), MADCAM-1 antisense probe (E), and GlyCAM-1 antisense riboprobe (F). The sections were then incubated with alkaline phosphatase-conjugated antidigoxigenin, and the digoxigenin-labeled compounds were detected as described in *Materials and Methods*. Bar, 100  $\mu$ m.

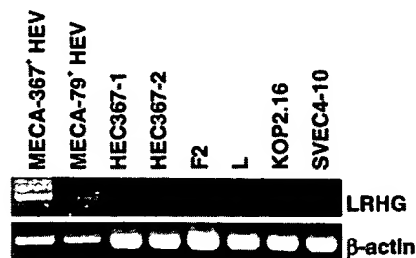
cDNA library. We then compared this gene expression profile with the profile we had obtained previously from peripheral LN HEV cells (23) and with expression profiles obtained from 35 tissues (see <http://bodymap.ims.u-tokyo.ac.jp/> for details). The appropriateness of the HEV cell preparation method we used was verified by the observations that MADCAM-1 was expressed only in the MECA-367<sup>+</sup> HEV cell preparation but not in the MECA-79<sup>+</sup> cell preparation and that GlyCAM-1 is expressed in the MECA-79<sup>+</sup> cell preparation but not in the MECA-367<sup>+</sup> cell preparation (Table II). These findings are fully compatible with the previous histological observation that these addressins are expressed in a tissue-specific manner (4).

Of 2101 sequences that we examined in the MECA-367<sup>+</sup> HEV cell library, 1304 were apparently derived from independent genes, and >60% were previously unreported, indicating that HEV cells express a large number of unique genes, a majority of which are as yet uncharacterized. Of the known genes, MECA-367<sup>+</sup> HEV cells expressed typical endothelial markers such as endoglin, ICAM-1, and L6 Ag, confirming their endothelial origin. MECA-367<sup>+</sup> HEV cells also expressed some genes in common with MECA-79<sup>+</sup> HEV cells but not CD31<sup>+</sup> endothelial cells; these common genes may represent HEV-specific genes or genes that are expressed in activated endothelial cells. These include the genes for SLC, KC, mac25, DARC, TAG7, ApoE, and transglutaminase. SLC is constitutively expressed in LN HEV cells (11) and mice deficient in SLC in lymphoid tissues (13) and those deficient in its receptor

CCR7 (44) show a selective defect in the migration of T cells and dendritic cells. DARC and mac25 expression in LN HEV cells was previously reported (23, 24). However, the expression of an inflammatory chemokine, KC, and an inflammation-related cytokine, TAG7, in HEV has not been reported previously. Both of these molecules attract and activate nonlymphoid-type inflammatory cells (36, 37), although these cell types do not migrate across HEVs under normal conditions. Currently, we do not know whether these molecules are expressed in HEVs at the protein level, but if so, there must be a mechanism whereby the function of these inflammatory mediators is abrogated. It is interesting that a putative scavenger for chemokines that binds KC, DARC (38), is expressed in HEV cells. ApoE also seems to be highly expressed in both MECA-367<sup>+</sup> HEV cells and MECA-79<sup>+</sup> HEV cells, but not in CD31<sup>+</sup> flat endothelial cells. Although ApoE is a ligand for several lipoprotein receptors and is known to play a major role in the hepatic clearance of remnant lipoproteins (45), it also stimulates the incorporation of <sup>35</sup>SO<sub>4</sub> and the production of heparan sulfate in endothelial cells (41). Because intensive incorporation of <sup>35</sup>SO<sub>4</sub> and production of heparan sulfate proteoglycans are observed in HEV cells in vivo (3), an increased expression of ApoE may be functionally related to the unique biosynthetic activity of HEV cells. Transglutaminase is another protein expressed in both types of HEV cells but not in CD31<sup>+</sup> flat endothelial cells. This protein is expressed in human endothelial cells, and its down-regulation leads to alterations in spreading and adhesion (46), although its biological significance remains unknown.

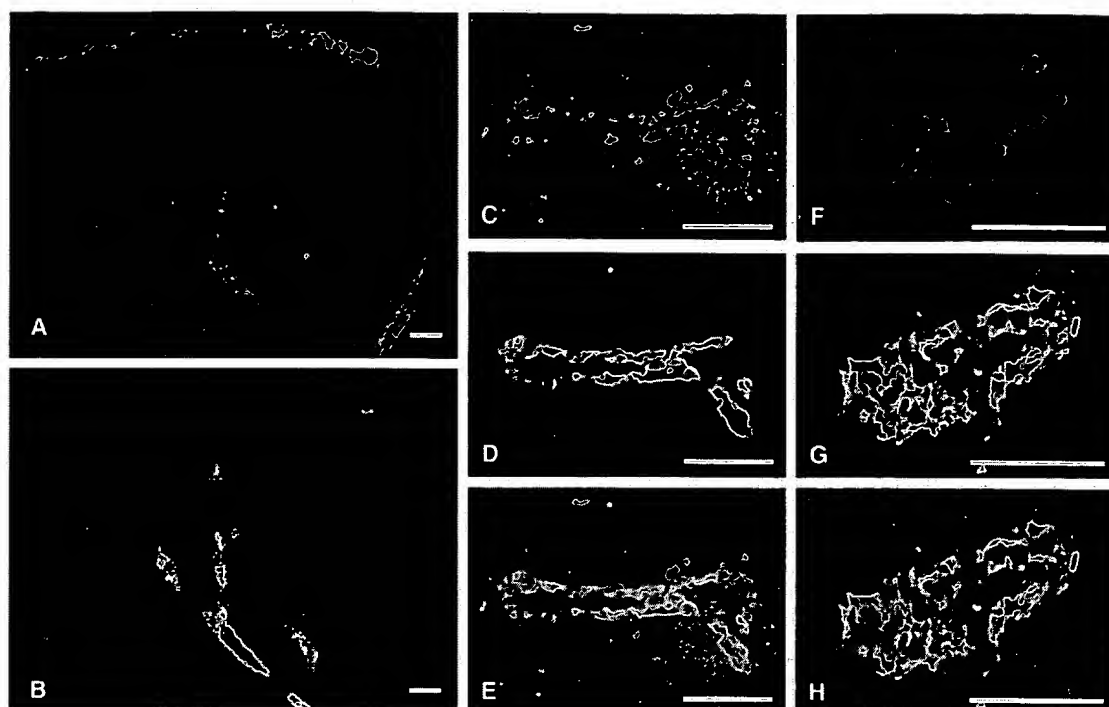
Several genes were found only in MECA-367<sup>+</sup> HEV cells, and not in MECA-79<sup>+</sup> HEV cells, normal endothelial cells, or T and B lymphocytes. They include those encoding MADCAM-1, VE-cadherin, lactadherin, Ly-6C.2, IL-2R  $\gamma$ -chain, SPI3, cyclin D1,  $\alpha$ -amylase-2, and MyD118. However, except for MADCAM-1, the transcripts of these genes appeared at such low to moderate frequencies that the significance of this observation is currently unclear. Also, we do not know whether the proteins encoded by these genes are selectively expressed in MECA-367<sup>+</sup> HEV cells. Further examination is required to verify the differential expression of these various molecules in HEV cells.

The present study demonstrated that a leucine-rich protein, LRHG, is an HEV marker in LNs and adds it to a growing list of novel molecules expressed preferentially in HEV. LRHG belongs to the LRR family; it bears eight LRR in tandem arrays and has an



**FIGURE 5.** RT-PCR analysis of LRHG expression in various cells. Agarose gel electrophoresis analysis of cDNA fragments amplified by PCR using LRHG-specific or  $\beta$ -actin primers. LRHG mRNA was detected in purified MECA-367<sup>+</sup> HEV cells and purified MECA-79<sup>+</sup> HEV cells but not in other cell lines (HEC367-1, HEC367-2, F2, KOP2.16, SVEC4-10, and L cells)



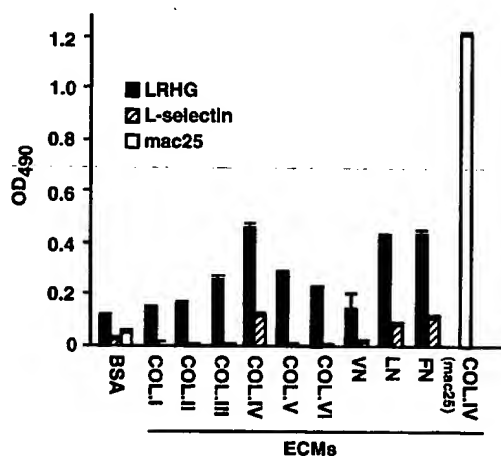


**FIGURE 6.** Immunohistochemical analysis of LRHG expression in mouse mesenteric LN. Double staining of mouse mesenteric LN cryosections with Abs against LRHG (A, C, and F) and MECA-367 (B and D) or MECA-79 (G). E and H show the combined images of C and D and F and G, respectively. The sections were fixed and incubated first with LRHG-specific Ab and then with biotin-conjugated anti-rabbit Ab, enhanced by ABC reagent, and stained by Vector Red (red). Next, the sections were incubated with FITC-conjugated MECA-367 or MECA-79 (green). White bar, 50  $\mu$ m.

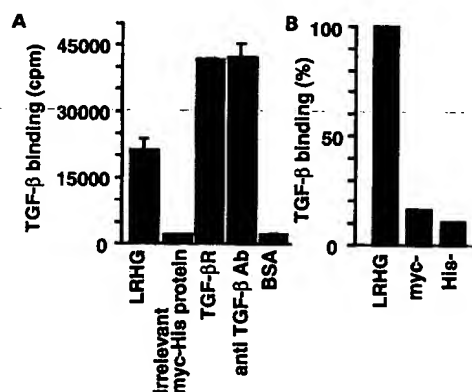
extremely similar domain structure to and high amino acid homology (67%) with human leucine-rich  $\alpha_2$ -glycoprotein, an LRR family member initially identified in human plasma (26). Judging from the extent of the homology, LRHG is likely to be a mouse homolog of this protein. The human leucine-rich  $\alpha_2$ -glycoprotein is a secretory protein with characteristics of an acute phase protein, in that its plasma level increases in the early stage of inflammation (47), but otherwise its function is unknown.

Although *in situ* hybridization analysis showed clearly that the mRNA expression of LRHG is restricted to HEV cells (Fig. 4, A

and C), immunohistochemical analysis with a polyclonal Ab showed LRHG staining in several layers of lymphocytes surrounding the HEVs as well as in HEV cells (Fig. 6). The staining in the lymphocytes became fainter the farther away they were from the HEVs (Fig. 6, C and F), indicating that the LRHG protein may be secreted from HEV cells where it binds lymphocytes immigrating into the LN from HEVs; it may then be lost from the surface of the lymphocytes as the lymphocytes migrate further into the LN cortex. Preliminary studies indicate that LRHG binds to a certain type of lymphoid cell, although its exact phenotype remains unclear (K. Saito and T. Tanaka, unpublished observation).



**FIGURE 7.** Binding of LRHG to various ECM proteins. Microtiter plates coated with various ECM proteins were incubated with myc-His-tagged LRHG, mac25, or L-selectin protein for 1 h at room temperature. The amount of bound myc-His-tagged protein was determined by ELISA as described in *Materials and Methods*. Data are shown with standard deviations of triplicate determinations. FN, Fibronectin; LN, laminin; VN, vitronectin; COL, collagen.



**FIGURE 8.** Binding of TGF- $\beta$  to rLRHG. A,  $^{125}$ I-labeled TGF- $\beta$  binding to immobilized rLRHG. rL-selectin was used as an irrelevant myc-His protein. Anti-TGF- $\beta$  mAb and rhTGF- $\beta$ R were used as positive controls. B, rLRHG was absorbed with anti-myc-agarose (myc-) or  $\text{Ni}^{2+}$ -charged agarose (His-) and immobilized on 96-well ELISA plates, and binding assays were performed. The binding is expressed as a percent of  $^{125}$ I-labeled TGF- $\beta$  bound to rLRHG.



Girard et al. (17) have identified an adhesion-regulating secretory protein, hevin, in HEVs, that has an antiadhesive effect on endothelial cells. Our preliminary studies indicate that LRHG does not have obvious antiadhesive properties (K. Saito and T. Tanaka, unpublished observation). Rather, LRHG binds to various ECM proteins such as fibronectin, laminin, and collagen that are abundant in the basal lamina of HEVs; hence, it may serve to mediate adhesion between HEV cells and the adjacent basal lamina.

LRHG appears to be a TGF- $\beta$ -binding LRR protein, similar to decorin and biglycan. The myc-His-tagged rLRHG bound TGF- $\beta$ . At present, we do not know whether TGF- $\beta$  binds to the LRR domain of LRHG. Nevertheless, the ability of LRHG to bind TGF- $\beta$  is interesting because TGF- $\beta$  is a cytokine that negatively regulates cell adhesion. TGF- $\beta$  inhibits lymphocyte adhesion to TNF- $\alpha$ -stimulated or IFN- $\gamma$ -stimulated endothelial cells (48). Although little is known about the expression of TGF $\beta$  in the LN paracortex, including in HEVs, it is interesting to speculate that LRHG localized to the HEV area serves as an anchoring molecule for TGF- $\beta$  or the like, thus helping the cytokine to form a concentration gradient around HEVs to regulate lymphocyte adhesiveness and migration in this area. Newly immigrating lymphocytes are likely to be adhesive to the parenchymal ECM, because their integrins have been recently activated by chemokines on the surface of HEVs (11). For these cells to successfully leave the HEV area to migrate further to the appropriate anatomical compartments, including T-dependent areas and follicular regions in the cortex, their adhesion to the ECM may have to be down-regulated by some mechanism(s).

Collectively, a variety of molecules that have been implicated in the regulation of cell adhesion are uniquely expressed in HEV cells. Further studies with LRHG and other molecules differentially expressed in HEV cells, and those molecules apparently expressed selectively in mesenteric but not peripheral LN HEV cells, may help elucidate the complex mechanism of tissue-specific lymphocyte trafficking across HEVs and the subsequent positioning of different lymphocyte subsets into various microcompartments in the LN.

## Acknowledgments

We thank Dr. E. C. Butcher for providing the mAb MECA-367 and MECA-79, Dr. T. Koga for the polyclonal anti-LRHG Abs, Dr. H. Kawashima for soluble L-selectin, and D. Nagakubo for myc-His-tagged mac25. We also thank Drs. H. Kawashima and T. Murai for stimulating discussion and critical reading of the manuscript.

## References

- Butcher, E. C., and L. J. Picker. 1996. Lymphocyte homing and homeostasis. *Nature* 372:60.
- Imhof, B. A., and D. Dunon. 1995. Leukocyte migration and adhesion. *Adv. Immunol.* 58:345.
- Kraal, G., and R. E. Mebius. 1996. High endothelial venules: lymphocyte traffic control and controlled traffic. *Adv. Immunol.* 65:347.
- Streeter, P. R., E. L. Berg, B. T. N. Rouse, R. F. Bargatze, and E. C. Butcher. 1988. A tissue-specific endothelial cell molecule involved in lymphocyte homing. *Nature* 331:41.
- Lasky, L. A., M. S. Singer, D. Dowbenko, Y. Imai, W. J. Henzel, C. Grimley, C. Fennie, N. Gillett, S. R. Watson, and S. D. Rosen. 1992. An endothelial ligand for L-selectin is a novel mucin-like molecule. *Cell* 69:927.
- Baumhuter, S., M. S. Singer, W. Henzel, S. Hemmerich, M. Renz, S. D. Rosen, and L. A. Lasky. 1993. Binding of L-selectin to the vascular sialomucin CD34. *Science* 262:436.
- Sassetti, C., K. Tangemann, M. S. Singer, D. B. Kershaw, and S. D. Rosen. 1998. Identification of podocalyxin-like protein as a high endothelial venule ligand for L-selectin: parallel to CD34. *J. Exp. Med.* 187:1965.
- Berg, E. C., L. M. McEvoy, C. Berlin, R. F. Bargatze, and E. C. Butcher. 1993. L-selectin-mediated lymphocyte rolling on MAdCAM-1. *Nature* 366:695.
- Bargatze, R., M. A. Jutila, and E. C. Butcher. 1995. Distinct Roles of L-selectin and integrin  $\alpha_4\beta_7$  and LFA-1 in lymphocyte homing to Peyer's patch-HEV in situ: the multistep model confirmed and refined. *Immunity* 3:99.
- Nagira, M., T. Imai, R. Yoshida, S. Takagi, M. Iwasaki, M. Baba, Y. Tabira, J. Akagi, H. Nomiyama, and O. Yoshie. 1998. A lymphocyte-specific CC chemokine, secondary lymphoid tissue chemokine (SLC), is a highly efficient chemoattractant for B cells and activated T cells. *Eur. J. Immunol.* 28:1516.
- Gunn, M. D., K. Tangemann, C. Tam, J. G. Cyster, S. D. Rosen, and L. T. Williams. 1998. A chemokine expressed in lymphoid high endothelial venules promotes the adhesion and chemotaxis of naive T lymphocytes. *Proc. Natl. Acad. Sci. USA* 95:258.
- Pachynski, R. K., S. W. Wu, M. D. Gunn, and D. J. Erle. 1998. Secondary lymphoid-tissue chemokine (SLC) stimulates integrin  $\alpha_4\beta_7$ -mediated adhesion of lymphocytes to mucosal addressin cell adhesion molecule-1 (MAdCAM-1) under flow. *J. Immunol.* 161:952.
- Gunn, M. D., S. Kyuwa, C. Tam, T. Kakiuchi, K. Matsuzawa, L. T. Williams, and H. Nakano. 1999. Mice lacking expression of secondary lymphoid organ chemokine have defects in lymphocyte homing and dendritic cell localization. *J. Exp. Med.* 189:451.
- Breitfeld, D., L. Ohl, E. Kremmer, J. Ellwart, F. Sallusto, M. Lipp, and R. Förster. 2000. Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *J. Exp. Med.* 192:1545.
- Schaerli, P., K. Willmann, A. B. Lang, M. Lipp, P. Loetscher, and B. Moser. 2000. CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function. *J. Exp. Med.* 192:1553.
- Springer, T. A. 1994. Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* 76:301.
- Girard, J.-P., and T. A. Springer. 1995. Cloning from purified high endothelial venule cells of hevin, a close relative of the antiadhesive extracellular matrix protein SPARC. *Immunity* 2:113.
- Hiraoka, N., B. Petryniak, J. Nakayama, S. Tsuboi, M. Suzuki, J.-C. Yeh, D. Izawa, T. Tanaka, M. Miyasaka, J. B. Lowe, and M. Fukuda. 1999. A novel, high endothelial venule-specific sulfotransferase express 6-sulfo sialyl Lewis<sup>x</sup>, an L-selectin ligand displayed by CD34. *Immunity* 11:79.
- Bistrup, A., S. Bhakta, J. K. Lee, Y. Y. Belov, M. D. Gunn, F.-R. Zuo, C.-C. Huang, R. Kannagi, S. D. Rosen, and S. Hemmerich. 1999. Sulfotransferases of two specificities function in the reconstitution of high endothelial cell ligands for L-selectin. *J. Cell Biol.* 145:899.
- Aurrand-Lions, M., L. Duncan, C. Ballestrem, and B. A. Imhof. 2001. JAM-2, a novel immunoglobulin superfamily molecule, expressed by endothelial and lymphatic cells. *J. Biol. Chem.* 276:2733.
- Palmeri, D., A. V. Zante, C.-C. Huang, S. Hemmerich, and S. D. Rosen. 2000. Vascular endothelial junction-associated molecule, a novel member of the immunoglobulin superfamily, is localized to intercellular boundaries of endothelial cells. *J. Biol. Chem.* 275:19139.
- Sassetti, C., A. V. Zante, and S. D. Rosen. 2000. Identification of endoglycan, a member of the CD34/podocalyxin family of sialomucins. *J. Biol. Chem.* 275:9001.
- Izawa, D., T. Tanaka, K. Saito, H. Ogihara, T. Usui, S. Kawamoto, K. Matsubara, K. Okubo, and M. Miyasaka. 1999. Expression profile of active genes in mouse lymph node high endothelial cells. *Int. Immunol.* 11:1989.
- Girard, J.-P., E. S. Baekkevold, T. Yamanaka, G. Haraldsen, P. Brandtzaeg, and F. Amalric. 1999. Heterogeneity of endothelial cells: the specialized phenotype of human endothelial venules characterized by suppression subtractive hybridization. *Am. J. Pathol.* 155:2043.
- Kobe, B., and J. Deisenhofer. 1994. The leucine-rich repeat: a versatile binding motif. *Trend. Biochem. Sci.* 19:415.
- Takahashi, N., Y. Takahashi, and F. W. Putnam. 1985. Periodicity of leucine and tandem repetition of a 24-amino acid segment in the primary structure of leucine-rich  $\alpha_2$ -glycoprotein of human serum. *Proc. Natl. Acad. Sci. USA* 82:1906.
- Streeter, P. R., B. T. N. Rouse, and E. C. Butcher. 1988. Immunohistologic and functional characterization of a vascular addressin involved in lymphocyte homing into peripheral lymph nodes. *J. Cell Biol.* 107:1853.
- Toyama-Sorimachi, N., K. Miyake, and M. Miyasaka. 1993. Activation of CD44 induces ICAM-1/LFA-1-independent,  $\text{Ca}^{2+}$ ,  $\text{Mg}^{2+}$ -independent adhesion pathway in lymphocyte-endothelial cell interaction. *Eur. J. Immunol.* 23:439.
- Toda, K., K. Tsujioka, Y. Maruguchi, K. Ishii, Y. Miyachi, K. Kuribayashi, and S. Imamura. 1990. Establishment and characterization of a tumorigenic murine vascular endothelial cell line (F-2). *Cancer Res.* 50:5526.
- O'Connell, K. A., and M. Edidin. 1990. A mouse lymphoid endothelial cell line immortalized by simian virus 40 binds lymphocytes and retains functional characteristics of normal endothelial cells. *J. Immunol.* 144:521.
- Ezaki, T., L. Yao, and K. Matsuno. 1995. The identification of proliferating cell nuclear antigen (PCNA) on rat tissue cryosections and its application to double immunostaining with other markers. *Arch. Histol. Cytol.* 58:103.
- Hering, T. M., J. Kollar, T. D. Huynh, and J. B. Varelans. 1996. Purification and characterization of decorin core protein expressed in *Escherichia coli* as a maltose-binding protein fusion. *Anal. Biochem.* 240:98.
- Okubo, K., N. Hori, R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* 2:173.
- Anderson, N. D., A. O. Anderson, and R. G. Wyllie. 1976. Specialized structure and metabolic activities of high endothelial venules in rat lymphatic tissues. *Immunology* 31:455.
- Barbara, N. P., J. L. Wrana, and M. Letarte. 1999. Endoglin is an accessory protein that interacts with the signaling receptor complex of multiple members of the transforming growth factor- $\beta$  superfamily. *J. Biol. Chem.* 274:584.
- Oquendo, P., J. Alberta, D. Z. Wen, J. L. Graycar, R. Derynck, and C. D. Stiles. 1989. The platelet-derived growth factor-inducible KC gene encodes a secretory protein related to platelet  $\alpha$ -granule protein. *J. Biol. Chem.* 264:4133.
- Kiselev, S. L., O. S. Kusukova, E. V. Korobko, E. B. Prokhortchouk, A. A. Kabishev, E. M. Lukandin, and G. P. Georgiev. 1998. Molecular cloning

- and characterization of the mouse *tag7* gene encoding a novel cytokine. *J. Biol. Chem.* 273:18633.
38. Luo, H., A. Chaudhuri, K. R. Johnson, K. Neote, V. Zbrzezna, Y. He, and A. O. Pogo. 1997. Cloning, characterization, and mapping of a murine promiscuous chemokine receptor gene: homolog of the human Duffy gene. *Genome Res.* 7:932.
39. Akaogi, K., Okabe, Y., Sato, J., Nagashima, Y., Yasumitsu, H., Sugahara, K., and Miyazaki, K. 1996. Specific accumulation of tumor-derived adhesion factor in tumor blood vessels and in capillary tube-like structures of cultured vascular endothelial cells. *Proc. Natl. Acad. Sci. USA* 93:8384.
40. Ritter, S. J., and P. J. A. Davies. 1998. Identification of a transforming growth factor- $\beta$ 1/bone morphogenic protein 4 (TGF- $\beta$ 1/BMP4) response element within the mouse tissue transglutaminase gene promoter. *J. Biol. Chem.* 273:12798.
41. Paka, L., Y. Kako, J. C. Obunike, and S. Pillarisetti. 1999. Apolipoprotein E containing high density lipoprotein stimulates endothelial production of heparan sulphate rich in biologically active heparin-like domains. *J. Biol. Chem.* 274:4816.
42. Iozzo, R. V. 1999. The biology of the small leucine-rich proteoglycans. *J. Biol. Chem.* 274:18843.
43. Yamaguchi, Y., D. M. Mann, and E. Ruoslahti. 1990. Negative regulation of transforming growth factor- $\beta$  by the proteoglycan decorin. *Nature* 346:281.
44. Förster, R., A. Schubel, D. Breitfeld, E. Kremmer, I. R.-Müller, E. Wolf, and M. Lipp. 1999. CCR7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs. *Cell* 99:23.
45. Ishibashi, S., S. Perry, Z. Chen, J. Osuga, M. Shimada, K. Ohashi, K. Harada, Y. Yazaki, and N. Yamada. 1996. Role of the low density lipoprotein (LDL) receptor pathway in the metabolism of chylomicron remnants: a quantitative study in knockout mice lacking the LDL receptor, apolipoprotein E, or both. *J. Biol. Chem.* 271:22422.
46. Jones, R. A., B. Nicholas, S. Mian, P. J. A. Davies, and M. Griffin. 1997. Reduced expression of tissue transglutaminase in a human endothelial cell line leads to changes in cell spreading, cell adhesion and reduced polymerisation of fibrinectin. *J. Cell Sci.* 110:2461.
47. Bini, L., B. Magi, B. Marzocchi, C. Cellesi, B. Berti, R. Raggiaschi, A. Rossolini, and V. Pallini. 1996. Two-dimensional electrophoretic patterns of acute-phase human serum proteins in the course of bacterial and viral diseases. *Electrophoresis* 17:612.
48. Chin, Y.-H., M.-W. Ye, J.-P. Cai, and X.-M. Xu. 1996. Differential regulation of tissue-specific lymph node high endothelial venule cell adhesion molecules by tumor necrosis factor and transforming growth factor- $\beta$ 1. *Immunology* 87:559.

**BLAST2 Search Results**

**EXHIBIT A**  
Docket No.: PF-0634 USN  
USSN: 09/831,455

Sequences Help

Retrieval BLAST2 FASTA ClustalW GCG Assembly Phrap Translation  
BLAST2 Manual

Confidential -- Property of Incyte Corporation SeqServer Version 4.6 Jan 2002

**Program: blastp**

**Sequence ID(s):**

☐ 1859618CD1 vs. genpept137

NCBI-BLASTP 2.0.10 [Aug-26-1999]



Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= 1859618CD1  
(347 letters)

Database: genpept137  
1,534,369 sequences; 474,463,515 total letters

Searching.....done

Sequences producing significant alignments:		Score (bits)	E Value
<input checked="" type="checkbox"/>	<u>g15321646</u> leucine-rich alpha-2-glycoprotein [Homo sapiens]	706	0.0
<input checked="" type="checkbox"/>	<u>g21707947</u> leucine-rich alpha-2-glycoprotein [Homo sapiens]	703	0.0
<input checked="" type="checkbox"/>	<u>g21315072</u> leucine-rich alpha-2-glycoprotein [Mus musculus]	419	e-116
<input checked="" type="checkbox"/>	<u>g18147610</u> leucine-rich HEV glycoprotein [Mus musculus]	419	e-116
<input checked="" type="checkbox"/>	<u>g15320546</u> leucine-rich alpha-2-glycoprotein [Mus musculus]	419	e-116
<input checked="" type="checkbox"/>	<u>g19911060</u> phospholipase A2 inhibitor [Elaphe quadrivirgata]	175	2e-42
<input checked="" type="checkbox"/>	<u>g3358089</u> phospholipase A2 inhibitor [Agkistrodon blomhoffii s	170	5e-41
<input checked="" type="checkbox"/>	<u>g19911058</u> phospholipase A2 inhibitor [Elaphe quadrivirgata]	169	8e-41
<input checked="" type="checkbox"/>	<u>g27769064</u> Similar to RIKEN cDNA 1300018K11 gene [Homo sapiens]	136	1e-30
<input checked="" type="checkbox"/>	<u>g21618740</u> Similar to RIKEN cDNA 1300018K11 gene [Homo sapiens]	136	1e-30

>g15321646 leucine-rich alpha-2-glycoprotein [Homo sapiens]  
Length = 347

Score = 706 bits (1803), Expect = 0.0  
Identities = 347/347 (100%), Positives = 347/347 (100%)

Query: 1 MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCPPA 60

```
Sbjct: 1      MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA 60
              MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA

Query: 61      EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV 120
              EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV
Sbjct: 61      EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV 120

Query: 121     LDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP 180
              LDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP
Sbjct: 121     LDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP 180

Query: 181     PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY 240
              PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY
Sbjct: 181     PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY 240

Query: 241     LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP 300
              LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP
Sbjct: 241     LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP 300

Query: 301     WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347
              WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ
Sbjct: 301     WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347
```

>g21707947 leucine-rich alpha-2-glycoprotein [Homo sapiens]  
Length = 347

Score = 703 bits (1795), Expect = 0.0  
Identities = 346/347 (99%), Positives = 346/347 (99%)

```
Query: 1      MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA 60
              MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA
Sbjct: 1      MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA 60

Query: 61      EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV 120
              EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV
Sbjct: 61      EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV 120

Query: 121     LDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP 180
              LDLTRNALTGLP GLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP
Sbjct: 121     LDLTRNALTGLPSGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLP 180

Query: 181     PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY 240
              PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY
Sbjct: 181     PGLLANFTLLRTLDDLGENQLETLPDLLRGPLQLERLHLEGKQLQVLGKDLLLPQPDRLRY 240

Query: 241     LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP 300
              LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP
Sbjct: 241     LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP 300

Query: 301     WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347
              WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ
Sbjct: 301     WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347
```

>g21315072 leucine-rich alpha-2-glycoprotein [Mus musculus]  
Length = 342

Score = 419 bits (1066), Expect = e-116  
Identities = 219/345 (63%), Positives = 254/345 (73%), Gaps = 7/345 (2%)

```
Query: 1      MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA 60
              M SW Q S ++ ++RTLFL LL G S K+C + +S GS++SC P
Sbjct: 1      MVSWQHQG--SLQDLKTCLARTLFLLLALL----GRVSSLKECLILQSAEGSTVSCHGPT 53
```

Query:	61	EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV	120
		E P LPADTVHL+VEF NLT LPA LQG L+ELHLSSN L++LSPE L PVP+LR	
Sbjct:	54	EFPSSLPADTVHLSVEFSNLTQLPAAALQGC PGLRELHLSSNRLQALSPELLAPVPRRA	113
Query:	121	LDLTRNALTG LPPGLFQASATLDTLV LKENQLEVLVSWLHGLKALGHLDLSGNRLRKLP	180
		LDLTRNAL LPPGLF SA L TLVL+ENQL + WL GL ALGHLDL+ N+L LP	
Sbjct:	114	LDLTRNALRSLPPGLFSTSANLSTLV LRENQLREVSAQWLQGLDALGHLDLAENQLSSLP	173
Query:	181	PGLLANFTLLR TLDLGENQLETLPD LLRGPLQLERLHLEG NKLQVLGKDLLLPQPDLRY	240
		GLLA+ L TLDLG N LE+LP LLRGP +L+RLHLEGN+LQ L LL PQP LR	
Sbjct:	174	SGLLASLGALHTLDLGYNLLES LPEGLLRGPRRLQRLHLEGNRLQRLED SLLAPQPFLRV	233
Query:	241	LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQP NWD MRDGFDISGNP	300
		LFLN N+L VA G+FQGL+ LDMLDLSNNSL+S P GLWA LG+P DM+DGFDIS NP	
Sbjct:	234	LFLNDNQLVG VATG SFQGLQHLDMLDLSNNSLSSTPPGLWAFLGRPTRDMQDGF DISHNP	293
Query:	301	WICDQNLS DLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAK	345
		WICD+NL+DL RWL A ++KMFSQNDTRCAGPEA+KGQ LL VA+	
Sbjct:	294	WICDKNLADLCRWLVANRNKMFSONDTRCAGPEAMKGORLLDVAE	338

>g18147610 leucine-rich HEV glycoprotein [Mus musculus]  
Length = 342

Score = 419 bits (1066), Expect = e-116  
Identities = 219/345 (63%), Positives = 254/345 (73%), Gaps = 7/345 (2%)

Query: 1	MSSWSRQRPKSPGGIQPHVSRTLFLLLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA	60
Sbjct: 1	M SW Q S ++ ++RTLFL LL G S K+C + +S GS++SC P	
Query: 61	EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV	120
Sbjct: 54	EFPSSLPADTVHLSVEFSNLTQLPAAALQGC PGLRELHLSSNRLQALSPELLAPVPRRA	113
Query: 121	LDLTRNALTG LPPGLFQASATLDTLV LKENQLEVLVSWLHGLKALGHLDLSGNRLRKLP	180
Sbjct: 114	LDLTRNAL LPPGLF SA L TLVL+ENQL + WL GL ALGHLDL+ N+L LP	173
Query: 181	PGLLANFTLLRTLDTLDGENQLETLPPDLLRGPLQLERLHLEGNKQLV LGKDLLLPQPDRLY	240
Sbjct: 174	SGLLASLGALHTLDLGYNNLES LPEGLLRGPRRLQRLHLEGNRLQRLED SLLAPQPFLRV	233
Query: 241	LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP	300
Sbjct: 234	LFLN N+L VA G+FQGL+ LDMLDLSNNSL+S P GLWA LG+P DM+DGFDIS NP	293
Query: 301	WICDQNLS DLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAK	345
Sbjct: 294	WICD+NL+DL RWL A ++KMFSQNDTRCAGPEA+KGQ LL VA+	338

```
>g15320546 leucine-rich alpha-2-glycoprotein [Mus musculus]
Length = 342
```

Score = 419 bits (1066), Expect = e-116  
Identities = 219/345 (63%), Positives = 254/345 (73%), Gaps = 7/345 (2%)

```
Query: 1    MSSWSRQRPKSPGGIOPHVSRTFLFLLLLAASAWGVTLSPKDCQVFRSDHGSSISCQPPA 60
           M SW  Q   S   ++  ++RTLFLL LL      G   S K+C + +S GS++SC  P
Sbjct: 1    MVSWQHQG--SLQDLKTCLARTFLFLLALL-----GRVSSLKECLILQSAEGSTVSCHGPT 53

Query: 61   EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRV 120
           E P  LPADTVHL+VEF NLT LPA  LQG   L+ELHLSSN L++LSPE L PVP+LR
Sbjct: 54   EFPSSLPADTVHLSVEFSNLTQLPAAALQGCPGLRELHLSSNRLOALSPELLAPVPRLRA 113
```

Query: 121 LDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLVSWLHGLKALGHLDLSGNRLRKLP 180  
LDLTRNAL LPPGLF SA L TLVL+ENQL + WL GL ALGHLDL+ N+L LP  
Sbjct: 114 LDLTRNALRSLPPGLFSTSANLSTLVLRNQLREVSAQWLQGLDALGHLDLAENQLSSLP 173

Query: 181 PGLLANFTLLRTLDTLGENQLETLPDPLLRGPLQLERLHLEGNKQLVGLKDLLLPQPDRLRY 240  
GLLA+ L TLDLG N LE+LP LLRG +L+RLHLEGN+LQ L LL PQP LR  
Sbjct: 174 SGLLASLGALHTLDLGYNLLESLEPEGLLRGPRRLQRLHLEGNRLQRLEDSLLAPQPFLRV 233

Query: 241 LFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFDISGNP 300  
LFLN N+L VA G+FQGL+ LDMLDLSNNSL+S P GLWA LG+P DM+DGFDIS NP  
Sbjct: 234 LFLNDNQLVGVATGSFQGLQHLMDLSDLSNNSLSTPPGLWAFGRPTRDMQDGFDISHNP 293

Query: 301 WICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAK 345  
WICD+NL+DL RWL A ++KMFSQNDTRCAGPEA+KGQ LL VA+  
Sbjct: 294 WICKNLADLCRWLVANRNKMFSQNDTRCAGPEAMKGQRLLDVAE 338

>g19911060 phospholipase A2 inhibitor [Elaphe quadrivirgata]  
Length = 332

Score = 175 bits (438), Expect = 2e-42

Identities = 102/294 (34%), Positives = 155/294 (52%), Gaps = 8/294 (2%)

Query: 56 CQPPA--EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSNNGLESLSPEFLR 113  
C P+ E P P T ++VEF ++ L LQG LQELHLS+N L++L R  
Sbjct: 41 CNSPSLHEFPTGFPVRTKLISVEFTQVSSLGVEALQGLPNLQELHLSNNRLKTLLSGLFR 100

Query: 114 PVPQLRVLDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLVSWLHGLKALGHLDLSG 173  
+P+L LDL+ N L LPP +F ++ +L L + EN+L L +SW LK L L L  
Sbjct: 101 NLPQLHTLDLSTNLLEDLPPEIFTSTTSLTLLSISENRLAKRLSWFETLKELRILSLDN 160

Query: 174 NRLRKLPPGLLANFTLLRTLDTLGENQLETLPDPLLRGPLQLERLHLEGNKQLVGLKDLLL 233  
N+L+++P L LDL N+L L PD+ G LERL LE N ++ +  
Sbjct: 161 NQLKEVPISCFDKLEKLTFLDLSSNRLHRLSPDMFSGLDNLERLSLENNPIRCIAPKSFH 220

Query: 234 PQPDRLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDG 293  
+P L + L L + G FQ L L +LDLS+N L L + P+ ++  
Sbjct: 221 GRPKLSIISLKNCSLTNIITGVFQPLNHLVLLDLSDELTM----LDPPVAIPSANL--S 274

Query: 294 FDISGNPWICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347  
D++GNPW C+ + +L W++ K ++S+ + CA P++ KG+ ++ +SQ  
Sbjct: 275 LDLTGNPWACNCRMDNLLTWVKEHKIDLYSKEEIVCAFPKSFKEEATSLHRSQ 328

>g3358089 phospholipase A2 inhibitor [Agkistrodon blomhoffii  
siniticus]  
Length = 331

Score = 170 bits (427), Expect = 5e-41

Identities = 105/294 (35%), Positives = 152/294 (50%), Gaps = 9/294 (3%)

Query: 56 CQPPA--EIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSNNGLESLSPEFLR 113  
C P+ E P PA ++VEF ++ L LQG LQELHLS+N L++L R  
Sbjct: 41 CNSPSLHEFPTGFPARAKMISVEFTQVSSLGVEALQGLPNLQELHLSNNRLKTLP SGLFR 100

Query: 114 PVPQLRVLDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLVSWLHGLKALGHLDLSG 173  
+PQL LDL+ N L LPP +F +++L L L ENQL L SW L L L L  
Sbjct: 101 NLPQLHTLDLSTNHLEDLPPEIFTNASSLILLPLSENQLAELHPSWFQTLGELRILGLDH 160

Query: 174 NRLRKLPPGLLANFTLLRTLDTLGENQLETLPDPLLRGPLQLERLHLEGNKQLVGLKDLLL 233  
N+++++P L +LDL N L L P++ G LE+L LE N +Q +  
Sbjct: 161 NQVKEIPISCFDKLKKLTSLDLSFNLLRRLAPEMFSGLDNLEKLILESNIQCIVGRTFH 220

Query: 234 PQPDRLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDG 293

P L L L + L + G FQ L QL++LDLS+N L ++ ++ + +  
 Sbjct: 221 WHPKLTVLSLKNSSLTNI-MGFFQPLEQLELLDLSDELTTMEPPVYKTSANLS----- 273  
 Query: 294 FDISGNPWICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347  
 D+SGNPW CD L +L W+ ++S+ + CA P+ KG+ ++ KSQ  
 Sbjct: 274 LDLSGNPWACDCRLDNLLTWVNEHNIHLYSKEEIVCASPKHFKGECATSLHKSQ 327

>g19911058 phospholipase A2 inhibitor [Elaphe quadrivirgata]  
 Length = 332

Score = 169 bits (425), Expect = 8e-41  
 Identities = 102/294 (34%), Positives = 153/294 (51%), Gaps = 8/294 (2%)

Query: 56 CQPPA--EIPGYLPADTVHlaveffnLTHLPANLLQGASKLQELHLSSNGLESLSPEFLR 113  
 C P+ E P P T ++VEF ++ L LQG LQELHLS+N L++L R  
 Sbjct: 41 CNSPSLHEFPTGFPVRTKLISVEFTQVSSLGVEALQGLPNLQELHLSNNRLKTLLSGLFR 100  
 Query: 114 PVPQLRVLDLTRNALTGLPPGLFQASATLDTLVLENQLEVLVSWLHGLKALGHLDLSG 173  
 +P+L+ LDL+ N L LPP +F + L L + EN+L L +SW LK L L L  
 Sbjct: 101 NLPELQTLDLSTNLLLEDLPPEIFANTTNLIQLSISENRLAELRLSWFETLKELTILGLDN 160  
 Query: 174 NRLRKLPPELLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKQLVLGKDLLL 233  
 N+L+++P L LDL N+L L PD+ G LERL LE N ++ +  
 Sbjct: 161 NQLKEIPISCFDKLKKLIFLDLSSNRLHRLSPDMFSGLDNLERLILEHNPIRCIAPKSFH 220  
 Query: 234 PQPDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDG 293  
 P L + L L + G FQ L L +LDLS+N L L + P+ ++  
 Sbjct: 221 GTPKLSIISLRNCSLTNIITGVFQPLNHLVLLDLSDELTM----LDPPVAIPSANL--S 274  
 Query: 294 FDISGNPWICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347  
 D++GNPW C+ + +L W++ K ++S+ + CA P++ KG+ ++ +SQ  
 Sbjct: 275 LDLTGNPWACNCRMDNLLTWVKEHKIDLYSKEEIVCAFPKSFKEEATSLHRSQ 328

>g27769064 Similar to RIKEN cDNA 1300018K11 gene [Homo sapiens]  
 Length = 560

Score = 136 bits (338), Expect = 1e-30  
 Identities = 99/292 (33%), Positives = 138/292 (46%), Gaps = 30/292 (10%)

Query: 80 LTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLPPGLFQAS 139  
 L LP L Q + L+ L+L+ N L L E P+ L+ L L+ NAL+GLP G+F  
 Sbjct: 172 LQALPRRLFQPLTHLKTNLNAQLPEELFHPLTSLQTLKLSNNALSGLPQGVFGKL 231  
 Query: 140 ATLDTLVLKEN-----QLEVLVSWLH-----GLKALGHLDLSGNR 175  
 +L L L N QL LE WL L L L L N  
 Sbjct: 232 GSLQELFLDSNNISELPQVFSQLFCLERLWLQRNAITHLPLSIFASLGNLTFLSLQWNM 291  
 Query: 176 LRKLPPELLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKQLVLGKDLLLPQ 235  
 LR LP GL A+ L L L NQLET+ L L L N + L +  
 Sbjct: 292 LRVLPAGLFAHTPCLVGLSLTHNQLETVAEGTFAHLSNLRSLMLSNAITHLPAGIFRDL 351  
 Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFD 295  
 +L L+L N L + FQ L +L++L LS N L ++PEG++ N+++ +  
 Sbjct: 352 EELVKLYLGSNNLTALHPALFQNLKLELLSLSKNQLTTLPEGIF----DTNYNLFN-LA 406  
 Query: 296 ISGNPWICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347  
 + GNPW CD +L+ L+ WLQ D++ + T CAGP +KGQ + A+ + Q  
 Sbjct: 407 LHGNPWQCDCHLAYLFNWLQOYTDRLLN-IQTYCAGPAYLKGQVVPALNEKQ 457

Score = 92.8 bits (227), Expect = 1e-17  
 Identities = 76/228 (33%), Positives = 102/228 (44%), Gaps = 1/228 (0%)

Query: 56 CQPPAEIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPV 115  
 CQ + G LP L V + +L N+ + L +L L+ N LE+L + +  
 Sbjct: 101 CQFRPDAFGGLPR-LEDLEVTGSSFLNLSTNIFSNLTSLGKLTNLFNMLEALPEGLFQHL 159

Query: 116 PQLRVLDLTRNALTGLPPGLFQASATLDTLVLENQLEVLVSWLHGLKALGHLDLGSNR 175  
 L L L N L LP LFQ L TL L +N L L H L +L L LS N  
 Sbjct: 160 AALESHLQGNRLQALPRRLFQPLTHLKTNLNAQNLLAQLPPELFHPLTSLQTLKLSNNA 219

Query: 176 LRKLPPGLLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKLVGLKDLLLPQ 235  
 L LP G+ L+ L L N + LPP + LERL L+ N + L +  
 Sbjct: 220 LSGLPQGVFGKLGSLQELFLDSNNISELPPQVFSQLFCLERLWLQRNAITHLPLSIFASL 279

Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASL 283  
 +L +L L N L + AG F L L L++N L +V EG +A L  
 Sbjct: 280 GNLTFLSLQWNMLRVLPAGLFAHTPCLVGLSLTHNQLETVAEGTFAHL 327

Score = 92.1 bits (225), Expect = 2e-17

Identities = 78/262 (29%), Positives = 117/262 (43%), Gaps = 11/262 (4%)

Query: 26 LLLLAASAWGVTLSPKDC---QVFRSDHGSSISCQPPAEIPGYLPADTVHLAVEFFNLTH 82  
 LLLLA A + DC +VF SD + A +P +P T ++ + T  
 Sbjct: 27 LLLLARPAQPCPMGC-DCFVQEVFCSD-----EELATVPLDIPPYTKNIIFVETSFTT 78

Query: 83 LPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLPPGLFQASATL 142  
 L L ++ + L P+ +P+L L++T ++ L +F +L  
 Sbjct: 79 LETRAFGSNPNLTQVFLNTQLCQFRPDAFGGLPRLEDLEVTGSSFLNLSTNIFSNLTSL 138

Query: 143 DTLVLKENQLEVLVSWLHGLKALGHLDLGSNRLRKLPPGLLANFTLLRTLDLGENQLET 202  
 L L N LE L L AL L L GNRL+ LP L T L+TL+L +N L  
 Sbjct: 139 GKLTNLFNMLEALPEGLFQHLAALESHLQGNRLQALPRRLFQPLTHLKTNLNAQNLLAQ 198

Query: 203 LPPDLLRGPLQLERLHLEGNKLVGLKDLLLPQPDRLRYLFLNGNKLARVAAGAFQGLRQL 262  
 LP +L L+ L L N L L + + L+ LFL+ N ++ + F L L  
 Sbjct: 199 LPEELFHPLTSLQTLKLSNNALSGLPQGVFGKLGSLQELFLDSNNISELPPQVFSQLFCL 258

Query: 263 DMLDLSNNSLASVPEGLWASLG 284  
 + L L N++ +P ++ASLG  
 Sbjct: 259 ERLWLQRNAITHLPLSIFASLG 280

Score = 88.5 bits (216), Expect = 2e-16

Identities = 66/201 (32%), Positives = 92/201 (44%), Gaps = 1/201 (0%)

Query: 116 PQLRVLDLTRNALTGLPPGLFQASATLDTLVLENQLEVLVSWLHGLKALGHLDLGSNR 175  
 P + + + T L F ++ L +V QL GL L L+++G+  
 Sbjct: 64 PYTKNIIFVETSFTTLETRAFGSNPNLTQVFLNTQLCQFRPDAFGGLPRLEDLEVTGSS 123

Query: 176 LRKLPPGLLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKLVGLKDLLLPQ 235  
 L + +N T L L L N LE LP L + LE LHL+GN+LQ L + L P  
 Sbjct: 124 FLNLSTNIFSNLTSLGKLTNLFNMLEALPEGLFQHLAALESHLQGNRLQALPRRLFQPL 183

Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFD 295  
 L+ L L N LA++ F L L L LSNN+L+ +P+G++ LG D +  
 Sbjct: 184 THLKTNLNAQNLLAQLPPELFHPLTSLQTLKLSNNALSGLPQGVFGKLGSLQELFLDSNN 243

Query: 296 ISGNPWICDQNLSDLYR-WLQ 315  
 IS P L L R WLQ  
 Sbjct: 244 ISELPPQVFSQLFCLERLWLQ 264

Score = 57.4 bits (136), Expect = 6e-07

Identities = 48/153 (31%), Positives = 65/153 (42%), Gaps = 24/153 (15%)

Query: 73 LAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLP 132



L++++ L LPA L L L L+ N LE+++ + LR L L+ NA+T LP  
Sbjct: 285 LSLQWNMLRVLPAFLFAHTPCLVGLSLTHNQLETVAEGTFAHLSNLRSLMLSNAITHLP 344  
Query: 133 PGLFQASATLDTLVLENQLEVLVSWLHGLKALGHLDLSGNRLRKLPPGLLANFTLLRT 192  
G+F+ L+ L L L N L L P L N + L  
Sbjct: 345 AGIFR-----DLEELVKLYLGSNNLTALHPALFQNLKLEL 380  
Query: 193 LDLGENQLETLPDLLRGPLQLERLHLEGNKLQ 225  
L L +NQL TLP + L L L GN Q  
Sbjct: 381 LSLSKNQLTTLPEGIFDTNYNLFNLALHGNPWQ 413

Score = 56.2 bits (133), Expect = 1e-06  
Identities = 28/78 (35%), Positives = 45/78 (56%)

Query: 73 LAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLP 132  
L + + +THLPA + + +L +L+L SN L +L P + + +L +L L++N LT LP  
Sbjct: 333 LMLSNAITHLPAGIFRDLEELVKLYLGSNNLTALHPALFQNLKLELLSLSKNQLTTLTP 392  
Query: 133 PGLFQASATLDTLVLEN 150  
G+F + L L L N  
Sbjct: 393 EGIFDTNYNLFNLALHGN 410

>g21618740 Similar to RIKEN cDNA 1300018K11 gene [Homo sapiens]  
Length = 557

Score = 136 bits (338), Expect = 1e-30  
Identities = 99/292 (33%), Positives = 138/292 (46%), Gaps = 30/292 (10%)

Query: 80 LTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLPPGLFQAS 139  
L LP L Q + L+ L+L+ N L L E P+ L+ L L+ NAL+GLP G+F  
Sbjct: 169 LQALPRRLFQPLTHLKTNLNAQNLLAQLPEELFHPLTSLQTLKLSNNALSGLPQGVFGKL 228  
Query: 140 ATLDTLVLKEN-----QLEVLVSWLH-----GLKALGHLDLSGNR 175  
+L L L N QL LE WL L L L L N  
Sbjct: 229 GSLQELFLDSNNISELPPQVFSQFLCFLERLWLQRNAITHLPLSIFASLGNLTFLSLQWNM 288  
Query: 176 LRKLPPGLLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKLVGLKDLLLPQ 235  
LR LP GL A+ L L L NQLET+ L L L N + L +  
Sbjct: 289 LRVLPAGLFAHTPCLVGLSLTHNQLETVAEGTFAHLSNLRSLMLSNAITHLPAGIFRDL 348  
Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFD 295  
+L L+L N L + FQ L +L++L LS N L ++PEG++ N+++ +  
Sbjct: 349 EELVKLYLGSNNLTALHPALFQNLKLELLSLSKNQLTTLPEGIF----DTNYNLFN-LA 403  
Query: 296 ISGNPWICDQNLSDLYRWLQAQKDKMFSQNDTRCAGPEAVKGQTLLAVAKSQ 347  
+ GNPW CD +L+ L+ WLQ D++ + T CAGP +KGQ + A+ + Q  
Sbjct: 404 LHGNPWQCDCHLAYLFWNLQYTDRLN-IQTYCAGPAYLKGQVVPALNEKQ 454

Score = 92.8 bits (227), Expect = 1e-17  
Identities = 76/228 (33%), Positives = 102/228 (44%), Gaps = 1/228 (0%)

Query: 56 CQPPAEIPGYLPADTVHLAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPV 115  
CQ + G LP L V + +L N+ + L +L L+ N LE+L + +  
Sbjct: 98 CQFRPDAFGGLPR-LEDLEVTGSSFLNLSTNIFSNLTSLGKLTNLFNMLEALPEGLFQHL 156  
Query: 116 PQLRVLDLTRNALTGLPPGLFQASATLDTLVLENQLEVLVSWLHGLKALGHLDLSGNR 175  
L L L N L LP LFQ L TL L +N L L H L +L L LS N  
Sbjct: 157 AALESHLQGNLQALPRRLFQPLTHLKTNLNAQNLLAQLPEELFHPLTSLQTLKLSNNA 216  
Query: 176 LRKLPPGLLANFTLLRTLDLGENQLETLPDLLRGPLQLERLHLEGNKLVGLKDLLLPQ 235  
L LP G+ L+ L L N + LPP + LERL L+ N + L +  
Sbjct: 217 LSGLPQGVFGKLSLQELFLDSNNISELPPQVFSQFLCFLERLWLQRNAITHLPLSIFASL 276

Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASL 283  
+L +L L N L + AG F L L L++N L +V EG +A L  
Sbjct: 277 GNLTFLSLQWNMLRVLPAGLFAHTPCLVGLSLTHNQLETVAEGTFAHL 324

Score = 90.5 bits (221), Expect = 6e-17  
Identities = 77/262 (29%), Positives = 117/262 (44%), Gaps = 11/262 (4%)

Query: 26 LLLLAASAWGVTLSPKDC---QVFRSDHGSSISCQPPAEIPGYLPADTVHLAVEFFNLTH 82  
LLLLA A + DC +VF SD + A +P +P T ++ + T  
Sbjct: 24 LLLLARPAQPCPMGC-DCFVQEVFCSD-----EELATVPLDIPPYTKNIIFVETSFTT 75

Query: 83 LPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLPPGLFQASATL 142  
L L ++ + L P+ +P+L L++T ++ L +F +L  
Sbjct: 76 LETRAFGSNPNLTKVVFLLNTQLCQFRPDFAFGGLPRLEDLEVTGSSFLNLSTNIFSNLTS 135

Query: 143 DTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLPPGLLANFTLLRTL DLGENQLET 202  
L L N LE L L AL L L GN+L+ LP L T L+TL+L +N L  
Sbjct: 136 GKLTNLFNMLEALPEGLFQHLAALES LHLQGNQLQALPRRLFQPLTHLKTNLNAQNLLAQ 195

Query: 203 LPPDLLRGPLQLERLHLEGNKLVGKDLLLPQPDRLYLFLNGNKLARVAAGAFQGLRQL 262  
LP +L L+ L L N L L + + L+ LFL+ N ++ + F L L  
Sbjct: 196 LPEELFHPLTSLQTLKLSNNALSGLPQGVFGKLGSLQELFLDSNNISELPPQVFSQLFCL 255

Query: 263 DMLDLSNNSLASVPEGLWASLG 284  
+ L L N++ +P ++ASLG  
Sbjct: 256 ERLWLQRNAITHLPLSIFASLG 277

Score = 88.2 bits (215), Expect = 3e-16  
Identities = 66/201 (32%), Positives = 92/201 (44%), Gaps = 1/201 (0%)

Query: 116 PQLRVLDLTRNALTGLPPGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNR 175  
P + + + T L F ++ L +V QL GL L L+++G+  
Sbjct: 61 PYTKNIIFVETSFTTLETRAFGSNPNLTKVVFLLNTQLCQFRPDFAFGGLPRLEDLEVTGSS 120

Query: 176 LRKLPPGLLANFTLLRTL DLGENQLETLPDLLRGPLQLERLHLEGNKLVGKDLLLPQ 235  
L + +N T L L L N LE LP L + LE LHL+GN+LQ L + L P  
Sbjct: 121 FLNLSTNIFSNLTSLGKLTNLFNMLEALPEGLFQHLAALES LHLQGNQLQALPRRLFQPL 180

Query: 236 PDLRYLFLNGNKLARVAAGAFQGLRQLDMLDLSNNSLASVPEGLWASLGQPNWDMRDGFD 295  
L+ L L N LA++ F L L L LSNN+L+ +P+G++ LG D +  
Sbjct: 181 THLKTNLNAQNLLAQ LPEELFHPLTSLQTLKLSNNALSGLPQGVFGKLGSLQELFLDSNN 240

Query: 296 ISGNPWICDQNLSDLYR-WLQ 315  
IS P L L R WLQ  
Sbjct: 241 ISELPPQVFSQLFCLERLWLQ 261

Score = 57.4 bits (136), Expect = 6e-07  
Identities = 48/153 (31%), Positives = 65/153 (42%), Gaps = 24/153 (15%)

Query: 73 LAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLP 132  
L++++ L LPA L L L L+ N LE+++ + LR L L+ NA+T LP  
Sbjct: 282 LSLQWNMLRVLPAGLFAHTPCLVGLSLTHNQLETVAEGTFAHLNLRSLMLSNAITHLP 341

Query: 133 PGLFQASATLDTLVLKENQLEVLEVSWLHGLKALGHLDLSGNRLRKLPPGLLANFTLLRT 192  
G+F+ L+ L L L N L L P L N + L  
Sbjct: 342 AGIFR-----DLEELVKLYLGSNNLTALHPALFQNL SKLEL 377

Query: 193 LDLGENQLETLPDLLRGPLQLERLHLEGNKLQ 225  
L L +NQL TLP + L L L GN Q  
Sbjct: 378 LSLSKNQLTTLP EGIFDTNYNLFNLALHGNPWQ 410

Score = 56.2 bits (133), Expect = 1e-06  
Identities = 28/78 (35%), Positives = 45/78 (56%)

Query: 73 LAVEFFNLTHLPANLLQGASKLQELHLSSNGLESLSPEFLRPVPQLRVLDLTRNALTGLP 132  
L + + +THLPA + + +L +L+L SN L +L P + + +L +L L++N LT LP  
Sbjct: 330 LMLSNAITHLPAGIFRDLEELVKLYLGSNLTLALHPALFQNLKLELLSLSKNQLTTLTP 389

Query: 133 PGLFQASATLDTLVLEN 150  
G+F + L L L N  
Sbjct: 390 EGIFDTNYNLFLNALHGN 407

Database: genpept137  
Posted date: Sep 11, 2003 11:22 AM  
Number of letters in database: 474,463,515  
Number of sequences in database: 1,534,369

Lambda	K	H
0.319	0.137	0.414

Gapped

Lambda	K	H
0.270	0.0470	0.230

Matrix: BLOSUM62  
Gap Penalties: Existence: 11, Extension: 1  
Number of Hits to DB: 342469585  
Number of Sequences: 1534369  
Number of extensions: 14811600  
Number of successful extensions: 62566  
Number of sequences better than 10.0: 3073  
Number of HSP's better than 10.0 without gapping: 1933  
Number of HSP's successfully gapped in prelim test: 1157  
Number of HSP's that attempted gapping in prelim test: 29774  
Number of HSP's gapped (non-prelim): 14807  
length of query: 347  
length of database: 474,463,515  
effective HSP length: 56  
effective length of query: 291  
effective length of database: 388,538,851  
effective search space: 113064805641  
effective search space used: 113064805641  
T: 11  
A: 40  
X1: 16 ( 7.4 bits)  
X2: 38 (14.8 bits)  
X3: 64 (24.9 bits)  
S1: 41 (21.7 bits)

Graphical Viewer...

---

Submit sequences to:

---

